Department of Information and Service Economy

Adapting image processing and clustering methods to productive efficiency analysis and benchmarking: A cross disciplinary approach

Xiaofeng Dai

…res the interdisciplinary applications of computational …
…larly, this thesis focuses on problems in productive efficiency a…
…e hardly approachable or solvable using conventional methods.
…ve efficiency analysis, null or zero values are often produced due to the wrong sk…
…rtosis of the inefficiency distribution as against the distributional assumption on the ineffi…
…rm. This thesis uses the deconvolution technique, which is traditionally used in image processin… for noise removal, to develop a fully non-parametric method for efficiency estimation. Publication I and Publication II are devoted to this topic, with focus being laid on the cross-sectional case and panel case, respectively. Through Monte-Carlo simulations and empirical applications to Finnish electricity distribution network data and Finnish banking data, the results show that the Richardson-Lucy blind deconvolution method is insensitive to the distributional assumptions, robust to the data noise levels and heteroscedasticity on efficiency estimation.

In benchmarking, which could be the next step of productive efficiency analysis, the 'best practice' target may not perform under the same operational environment with the DMU under study. This would render the benchmarks impractical to follow and, consequently, adversely affects the managers to make the …orrect decisions on performance improvement of a DMU. This dissertation proposes a clustering-base… …hmarking framework in Publication III. In this framework, we group the DMUs into segmen… …ring methods based on certain metrics under interest, and estimate the efficiencies … …he segment-specific benchmark for DMUs within each cluster. The empi… … distribution network reveals that the proposed framework novels … …the differences of the operational environment among DM… …ng and efficiency estimation techniques are user… …We conducted a comparison analysis on t…

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 134/2016

Adapting image processing and clustering methods to productive efficiency analysis and benchmarking: A cross disciplinary approach

Xiaofeng Dai

Aalto University
School of Business
Department of Information and Service Economy
Management Science

Main dissertation advisor
Professor Timo Kuosmanen, Aalto University, Finland

Opponent
Professor Andrew Johnson, Texas A&M University, USA

Aalto University publication series
DOCTORAL DISSERTATIONS 134/2016

© Xiaofeng Dai

ISBN 978-952-60-6906-7 (printed)
ISBN 978-952-60-6907-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6907-4

Unigrafia Oy
Helsinki 2016

Finland

441 697
Printed matter

Author
Xiaofeng Dai

Name of the doctoral dissertation
Adapting image processing and clustering methods to productive efficiency analysis and benchmarking: A cross disciplinary approach

Publisher School of Business

Unit Management Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 134/2016

Field of research Quantitative Methods in Economics

Date of the defence 26 August 2016

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

This dissertation explores the interdisciplinary applications of computational methods in quantitative economics. Particularly, this thesis focuses on problems in productive efficiency analysis and benchmarking that are hardly approachable or solvable using conventional methods.

In productive efficiency analysis, null or zero values are often produced due to the wrong skewness or low kurtosis of the inefficiency distribution as against the distributional assumption on the inefficiency term. This thesis uses the deconvolution technique, which is traditionally used in image processing for noise removal, to develop a fully non-parametric method for efficiency estimation. Publications 1 and 2 are devoted to this topic, with focus being laid on the cross-sectional case and panel case, respectively. Through Monte-Carlo simulations and empirical applications to Finnish electricity distribution network data and Finnish banking data, the results show that the Richardson-Lucy blind deconvolution method is insensitive to the distributio-nal assumptions, robust to the data noise levels and heteroscedasticity on efficiency estimation.

In benchmarking, which could be the next step of productive efficiency analysis, the 'best practice' target may not perform under the same operational environment with the DMU under study. This would render the benchmarks impractical to follow and adversely affects the managers to make the correct decisions on performance improvement of a DMU. This dissertation proposes a clustering-based benchmarking framework in Publication 3. The empirical study on Finnish electricity distribution network reveals that the proposed framework novels not only in its consideration on the differences of the operational environment among DMUs, but also its extreme flexibility. We conducted a comparison analysis on the different combinations of the clustering and efficiency estimation techniques using computational simulations and empirical applications to Finnish electricity distribution network data, based on which Publication 4 specifies an efficient combination for benchmarking in energy regulation.

This dissertation endeavors to solve problems in quantitative economics using interdisciplinary approaches. The methods developed benefit this field and the way how we approach the problems open a new perspective.

Keywords Productive efficiency analysis, Benchmarking, Deconvolution, Clustering, StoNED

ISBN (printed) 978-952-60-6906-7	**ISBN (pdf)** 978-952-60-6907-4	
ISSN-L 1799-4934	**ISSN (printed)** 1799-4934	**ISSN (pdf)** 1799-4942
Location of publisher Helsinki	**Location of printing** Helsinki	**Year** 2016
Pages 89	**urn** http://urn.fi/URN:ISBN:978-952-60-6907-4	

Preface

Aalto University

The work presented in this thesis was carried out under the supervision of Prof. Timo Kuosmanen at the Department of Information and Service Economy at Aalto University during 08/2011 to 08/2015.

I would like to address my deepest gratitude to Prof. Timo Kuosmanen for introducing me these cutting-edge research topics in Quantitative Economics, encouraging me to work on interdisciplinary fields, and guiding me through all these projects.

I would like to express my warmest acknowledgement wholeheartedly to Dr. Juha Eskelinen for offering me finnish banking data.

I owe my great thankfulness to all the other group members of Prof. Timo Kuosmanen for various help and advices on work and daily life.

In addition, I am extremely grateful to my two reviewers, Dr. Andrew L. Johnson and Dr. Christopher F. Parmeter, for their pertinent and constructive comments given to my thesis and the fruitful discussions which help me improve my understandings towards quantitative methods in economics.

As the second doctoral degree I pursue, which was conducted while working as a post-doc in Cancer Genetics at Helsinki University Central Hospital and later an associate professor in Jiangnan University, China, I was extremely over-loaded and had gone through many indecisive moments regarding the continuity of this degree. It is my family and friends, no matter near or far, live or dead, that have encouraged me to finish this journey. I will never forget the delightful moments we spent together, the comprehensive mutual understandings we reached, and the ceaseless supports and blessings I received, which make my life full of joys, less confused and more encouraged. To them, I have nothing but gratefulness and would like to devote my sincere gratitude with all my heart.

For all the people that have helped or supported me during my doctoral studies, I dedicate this thesis to them and may they be blessed.

Wuxi, China, July 4, 2016,

Xiaofeng Dai

Contents

Preface **1**

Contents **3**

List of Publications **5**

Author's Contribution **7**

1. Background **11**

1.1 Background 11

1.1.1 Efficiency estimation 11

1.1.2 Benchmarking 11

1.2 Objectives 12

2. Methodologies **15**

2.1 Productive efficiency analysis and deconvolution 15

2.1.1 Productive efficiency analysis 15

2.1.2 Deconvolution 17

2.2 Benchmarking and clustering 26

2.2.1 Benchmarking 26

2.2.2 Clustering 26

3. Summary of the articles **33**

4. Concluding remarks **35**

Bibliography **37**

Publications **45**

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Xiaofeng Dai. Non-parametric efficiency estimation using Richardson-Lucy blind deconvolution. *European Journal of Operational Research*, 248, 731-739, 2016.

II Xiaofeng Dai. Non-parametric efficiency estimation in a panel setting: corrected Richardson-Lucy blind deconvolution. *Proceedings of 2015 International Conference on Management Engineering and Information Technology Application*, Hong Kong, China, April 19-20, 2015.

III Xiaofeng Dai, Timo Kuosmanen. Best-practice benchmarking using clustering methods: Application to energy regulation. *Omega*, 42 (1), 179-188, 2014.

IV Xiaofeng Dai. NMM-StoNED: a normal mixture model based stochastic semi-parametric benchmarking method. *International Journal of Business and Management Study*, 2 (2), 2015.

Author's Contribution

Publication I: "Non-parametric efficiency estimation using Richardson-Lucy blind deconvolution"

The author designed the study, carried out the analysis and drafted the manuscript.

Publication II: "Non-parametric efficiency estimation in a panel setting: corrected Richardson-Lucy blind deconvolution"

The author designed the study, carried out the analysis and drafted the manuscript.

Publication III: "Best-practice benchmarking using clustering methods: Application to energy regulation"

The author co-designed the study with Timo Kuosmanen, developed the clustering based benchmarking method, applied it to Finnish electricity distribution networks, and drafted the manuscript.

Publication IV: "NMM-StoNED: a normal mixture model based stochastic semi-parametric benchmarking method"

The author designed the study, carried out the analysis and drafted the manuscript.

Abbreviations

AIC: Akaike information criterion

AIC3: modified Akaike information criterion

ARMA: autoregressive moving average

AR: autoregressive

BIC: Bayesian information criterion

CNLS: convex nonparametric least squares

COLS: corrected ordinary least squares

CPI: consumer price index

cRLb: corrected Richardson-Lucy blind deconvolution

DEA: data envelopment analysis

DMU: decision making unit

EM: expectation maximization

EMA: energy market authority

FIR: finite impulse response

GDP: gross domestic product

GO: gene ontology

HOS: nonparametric methods based on high order statistics

HS: Hall and Simar

MA: moving average

MED: minimum entropy deconvolution

MM: method of moments

MSE: mean squared error

NAS-RIF: non-negativity and support constraints recursive inverse filtering

NMM: normal mixture model

IBD: iterative blind deconvolution

ICL-BIC: integrated classification likelihood-BIC

OLS: ordinary least squares

PP: parametric programming

RL: Richardson-Lucy

RLb: Richardson-Lucy blind deconvolution

SA: simulated annealing

SFA: stochastic frontier analysis

SOM: self organizing map

developing SS: fixed effects approach

StoNED: stochastic non-smooth envelopment of data

1. Background

1.1 Background

1.1.1 Efficiency estimation

According to traditional economics theory, all decision-making units (DMUs) operate efficiently. That is, DMUs produce the maximum output from given inputs at the lowest cost, which in terms of production function, means maximizing their productivity [1]. Thus, by superficial interpretation, the traditional economics theory means that no DMU is technically inefficient as they would be driven out of the market otherwise. However, this does not comply with what we observe in reality. Persistent efficiency differences are pointed out to exist virtually in all types of industries. By examining the determinants of efficiency differences, Syverson pointed out that the internal differences are determined by factors such as managerial talent and R&D, and the external divergence refers to the market conditions or operational environment [2].

The conventional approach in production or cost function estimation uses the linear regression methods without explicitly acknowledging the presence of technical inefficiency. Though conventional empirical models do allow deviations from the optimal production, these models usually under-estimate these deviations by taking them solely as a statistical error (see the discussion in [3]). In other words, the resulting residuals are considered as the estimation error and the interest focuses on studying the parameters of the production function itself. In cases where the residual is interested in, it is collapsed as a single productivity measure (see e.g. [2, 4], and the study focuses on the factors explaining variations in this residual but not its magnitude. Thus, it is impractical to decompose the technical inefficiency part from the residuals through conventional modeling. However, the efficiency needs to be correctly quantified for the managers to practically evaluate the productivity level of the DMUs and make the corresponding managerial decisions. Though approaches allowing explicit modeling of the inefficiencies are developed correspondingly, issues arisen from discordant distribution assumption on the inefficiency term have never been bypassed. This has led to the first objective of this thesis, i.e., developing a fully non-parametric efficiency estimation method.

1.1.2 Benchmarking

Besides efficiency itself, managers also need to be aware of the benchmarking target to make practical judgements on firm performance. Thus, the operation of DMUs should be compared with the ideal tech-

nology outputting the optimal amount of production. However, such theoretical scenarios hardly exist, and the technology has to be estimated from the observed data and compared with the best observed practices, namely the benchmark.

Benchmarking, by definition, is the process of comparing the performance metrics of a DMU to the best practices among all the DMUs. The management often identifies the best DMUs in their industry, or in another industry where similar processes exist, and compares the results and processes of those targets with their own. By doing so, they learn how well the targets perform and figure out why these targets are successful. This allows organizations to develop strategies on how to make improvements or adapt specific best practices to improve certain aspects of the performance. Though, benchmarking may be a one-off event, it is a continuous process where DMUs continually seek to improve their practices.

Among various benchmarking methods, DEA has long been used as a standard and important tool. The standard DEA assumes that all DMUs operate in a relatively similar environment [5] which, however, is not the case in practice. As the DMUs may seem inefficient given their poor environment, which is not actually caused by technical deficiency, it is intuitive that the comparison is meaningful only when the DMUs operate in a relatively similar environment. One major extension for all frontier methods including DEA on efficiency estimation is to account for the heterogeneity of the operational environment, which peels off the frontier in a sequential fashion to group DMUs into classes at different efficiency levels. However, these methods still could not take segment-wise differences into account, leading to the second objective of this thesis, i.e., developing strategies encompassing environmental divergence for benchmarking.

1.2 Objectives

Given the aforementioned background, the objectives of this thesis could be summarized as below:

- Developing a fully non-parametric inefficiency estimation method to 1) improve prediction accuracy and 2) resolve problems arisen from discordant parametric assumption on the inefficiency term which are unsolvable using conventional approach

- Developing strategies to output benchmarks that function in the same operational environment as the DMUs.

With the aforementioned objectives, this thesis focuses on interdisciplinary methods in achieving these goals. The deconvolution technique, conventionally applied in image processing for noise decomposition, was used for inefficiency estimation (objective 1), and the clustering method, traditionally used in

biology for gene classification, was adopted for benchmarking (objective 2).

2. Methodologies

2.1 Productive efficiency analysis and deconvolution

2.1.1 Productive efficiency analysis

Productive efficiency analysis, analyzing the productive efficiency of the units under study, is a classic problem in, e.g., economics, econometrics and statistics [11]. It is comprised of two parts, i.e., frontier estimation and error decomposition. Two approaches dominate this field, which are data envelopment analysis (DEA) [6,7] and stochastic frontier analysis (SFA) [8,9]. DEA is a static nonparametric method, which does not assume any particular functional form of the frontier but relies on the general regularity properties such as free disposability, convexity and assumptions concerning the returns to scale. This method, though values in its non-parametric form in frontier estimation, attributes all deviations from the frontier to the inefficiency, i.e., ignoring any stochastic noise in the data. SFA, on the other hand, is a parametric regression model, which requires exquisite specification of the functional form of the frontier. As rarely a specific functional form is justifiable by the economic theory, the flexible functional forms such as the translog or generalized McFadden are frequently used, which often violate the monotonicity, concavity/convexity, homogeneity conditions and sacrifice the flexibility [10]. However, SFA adopts a stochastic framework in its treatment of the deviation from the frontier, where the error term is decomposed into a non-negative inefficiency term and a random disturbance term comprising of random noise and measurement errors. Thus, the virtues of these two approaches complement each other, with DEA being nonparametric in frontier estimation and SFA being stochastic in error decomposition.

Many studies have considered DEA and SFA as competing alternatives. There has long been a lively debate on their relative pros and cons against each other which, though, tends to gain neutral tones in recent years, has led to the development of extensions of these approaches to account for their defects [11]. Though both methods have significantly evolved from their original forms, neither one clearly wins, and comparisons over the years only identify different circumstances where each method outperforms [12–14].

Efforts on bridging the gap between DEA and SFA have never been stopped ever since 1990s. Many success stories have been stemmed from the SFA side. Through replacing the parametric frontier function by a nonparametric specification estimable using techniques such as kernel regression or local maximum likelihood, semi-nonparametric stochastic methods were derived. Pioneer studies belonging to this branch include Fan's work [15] in the cross-sectional case and the research conducted by Kneip and

Simar [16] in treating panel data. This set of work employs kernel regression in frontier estimation while keeping the stochasticity of the SFA part in error decomposition. The distributional assumptions in these studies are imposed the same way as in SFA when decomposing the conditional expected inefficiency term from the residuals in the cross-sectional case [15], and are avoided in the panel case by making use of the information buried in such data [16]. Another set of work include Kumbhakar et al. [17] and Simar and Zelenyuk [18], which adopts the maximum likelihood method in frontier estimation while parameterizing the model in a similar way as the standard SFA. All model parameters are approximated by local polynomials in [17] and extended to multi-output technologies in [18]. In addition, monotonicity and concavity are imposed by applying DEA to the fitted values of Kumbhakar's model in [18].

From the DEA side, Banker and Maindiratta [19], in 1992, considered estimating the stochastic frontier model using maximum likelihood, subject to the global free disposability and convexity axioms adopted from DEA. This method combines the valuable features of both the classic DEA and SFA models whose resulting maximum likelihood problem is, however, technically impractical to solve. This bottleneck has not been solved until 2008, when theoretical links between DEA and the regression techniques were revealed [20, 21]. It is formally shown that DEA can be understood as a constrained special case of nonparametric least squares subject to shape constraints [20, 21]. Specifically, the classic output-oriented DEA estimator can be computed in the single-output case by solving the convex nonparametric least squares (CNLS) problem subject to monotonicity and concavity constraints characterizing the frontier and a sign constraint on the regression residuals [21]. Thus, Kuosmanen et al. proposed a method that estimates the model frontier shape using CNLS regression and developed a new two-stage method, namely stochastic non-smooth envelopment of data (StoNED) [21]. This approach does not assume any *a prior* functional form for the regression function. The classic DEA and SFA are both constrained special cases of this encompassing semiparametric frontier model, assuming that the observed data deviates from a non-parametric, DEA-style piecewise linear frontier production function due to a stochastic SFA-style composite error term, and such an error term is composed of homoscedastic noise and inefficiencies [22]. In the first stage, CNLS identifies the function best fitting the data from the family of continuous, monotonic increasing, concave functions that can be non-differentiable. In the second stage, the variance parameters of the stochastic inefficiency and noise terms are estimated based on the skewness of the CNLS residuals. The skewness of the residuals is attributed to the inefficiency term assuming that the noise term is symmetric. The variance parameters can be estimated by techniques such as the method of moments (MM) [8] and pseudolikelihood [15], provided with the parametric distributional assumptions of the inefficiency and the noise terms. In the cross-sectional setting, the distributional assumption is indispensable in distinguishing the inefficiency term from the noise. The time-invariant inefficiency components can be estimated in a fully nonparametric fashion by the standard fixed effects treatment analogous to the method proposed by Schmidt and Sickles [23] in the panel setting.

StoNED differs from the parametric or semi/nonparametric SFA in that it does not impose any assumptions on the functional form or smoothness, but builds on the global shape constraints which are equivalent to the free disposability and convexity axioms of DEA. On the other hand, it differs from DEA in its probabilistic treatment of the composite error term employing the entire observations for frontier estimation without being biased by outliers and noise. Given the advantages of StoNED as compared with DEA and SFA [24], it has been considered as the most efficient semi-parametric stochastic model in production efficiency analysis.

Since late 1970s, the field of productive efficiency analysis has undergone a plethora of empirical applications of DEA, SFA, and recently StoNED. Over the years, the applications of these methods have ranged from the micro to the aggregate macro level. Fried et al. (2008) identified around 50 different application areas of these methods [11], including accounting, advertising, auditing, law firms, airports, air transport, bank branches, bankruptcy prediction, benefit-cost analysis, community and rural health care, correctional facilities, credit risk evaluation, dentistry, discrimination, primary, secondary and tertiary education, elections, electricity distribution, electricity generation, macro and micro environmental applications, financial statement analysis, fishing, forestry, gas distribution, hospitals, hotels, inequality and poverty insurance, internet commerce, labor markets, libraries, location, macroeconomics, mergers, military, municipal services, museums, nursing homes, physicians and physician practices, police, ports, postal services, public infrastructure, rail transport, real estate investment trusts, refuse collection and recycling, sports, stocks, mutual funds, hedge funds, tax administration, telecommunications, urban transit, water distribution, world health organization. Productive efficiency analysis is the first step of benchmarking, i.e., the efficiencies estimated from productive efficiency analysis could be further used for identifying the efficient DMUs in benchmarking. In this thesis, we particularly focus on applications of efficiency analysis and benchmarking approaches developed on top of StoNED in bank branches [25–28] and electricity distribution [29–31], given data availability and their wide-applications in DMU incentivization.

2.1.2 Deconvolution

Deconvolution is a common technique traditionally applied for noise clearance in image processing. The original or true image is the ideal representation of the observed scene. In other words, the observation process is never perfect, i.e., uncertainties exist in the measurements which occur as blur, noise and other degradations in the recorded images. Thus, correct removal of these uncertainties has long been an important problem in image restoration. Classical approaches used for this purpose seek an estimate of the true image assuming the blur is known. Blind deconvolution, in contrast, tackles the much more difficult but realistic scenario where the degradation process is unknown. In general, the degradation is

nonlinear and spatially varying. However, it is assumed that the observed image is the output of a linear spatially invariant (LSI) system where noise is added. Thus, the problem becomes a blind deconvolution problem, with the unknown blur represented as a point spread function (PSF). The concepts involved and methods commonly used in deconvolution, and particularly blind deconvolution, are described in detail below.

Deconvolution concepts

Deconvolution is the process of estimating the clean original image from the corrupt noisy image as illustrated in Figure 2.1. It is a reverse operation of convolution which is a mathematical way of combining two signals to form a third signal. Similar with multiplication, addition and integration, convolution is a formal mathematical operation. Addition takes two numbers and produces a third number, while convolution takes two signals and produces a third signal. Such a notion is illustrated using a linear system in Figures 2.2 and 2.3, where the impulse response is called PSF in image processing. Expressed in words, the input signal convolved with PSF is equal to the output signal.

Figure 2.1. Images before and after deconvolution.

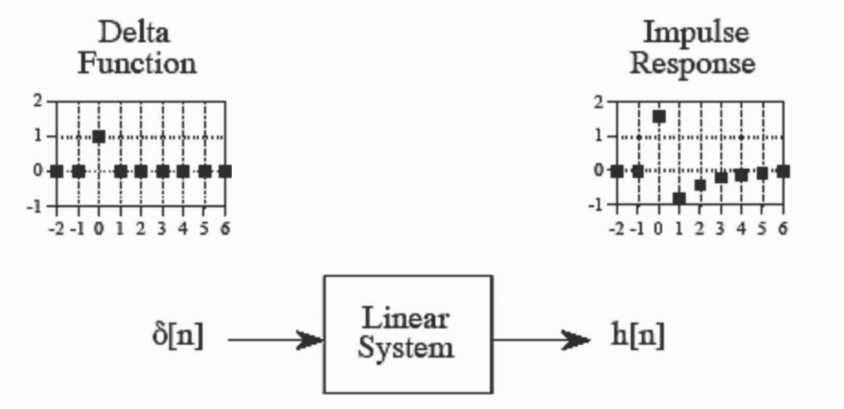


Figure 2.2. Illustration of PSF using a linear system. The delta function is a normalised impulse, which is identified by the Greek letter delta, $\delta[n]$; the PSF of a linear system is denoted by $h[n]$.

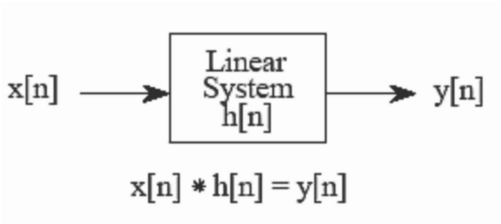


Figure 2.3. Illustration of convolution using a linear system. $x[n]$ and $y[n]$ are the input and output, respectively, of the linear system, where $h[n]$ is the PSF.

The PSF, by definition, is a function describing the response of an imaging system to a point object [32]. It describes the reaction of a dynamic system in response to some external changes as a function of time. In the context of economics, impulse response functions are usually called 'shocks' and used to model the reaction of economy in response to exogenous or endogenous impulses over time. Exogenous impulses include, e.g., changes in fiscal policy parameters such as government spending and tax rates, monetary policy parameters such as monetary base, technological parameters such as productivity, and preferences such as degree of impatience. Endogenous variables include, e.g., output, consumption, investment and employment at the time of shock and over subsequent points in time [33, 34]. For example, PSF can be used to model the impulse response of gross domestic product (GDP) growth rate to consumer price index (CPI). If the AR model of GDP is written as $y_t = \mu + \varepsilon_t + \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \cdots$, where y and x represent GDP and CPI, respectively, Φ_i can be interpreted as the response of GDP at

time t, i.e., y_t, to the one unit change of CPI at time $(t-i)$, i.e., x_{t-i}, given $\Phi_i = \frac{\partial y_t}{\partial x_{t-i}}$.

To further understand the mathematical background of convolution and deconvolution, two concepts are indispensable to introduce, i.e., the time domain and the frequency domain. These domains are used for analyzing mathematical functions with respect to time and frequency, respectively. A time domain graph shows how a signal changes over time, whereas a frequency domain graph illustrates how much of a signal lies within each given frequency band over a range of frequencies. A given function can be converted between the time and frequency domains by a pair of mathematical operators, i.e., Fourier transform and its inverse operation. The Fourier transform decomposes a function into the sum of an infinite number of sine wave frequency components, and the inverse Fourier transform converts the frequency domain function back to a time function. Thus, convolution is an operation in the time domain showing the multiplicative operation at the frequency domain. If $f(x)$ and $h(x)$ are the integrable functions with Fourier transforms $F(\omega)$ and $H(\omega)$, then $G(\omega) = F(\omega) \cdot H(\omega)$ in the frequency domain is equivalent to $g(x) = \int_{-\infty}^{+\infty} f(\tau)h(x-\tau)d_\tau = f(x) \otimes h(x)$ in the time domain, where $\otimes$ is the symbol denoting convolution [35]. Convolution can be viewed as the integral of the product of the two functions after one is reversed and shifted. Or, one can assume a sliding window which slides from $-\infty$ to $+\infty$, and convolution is a weighted average of function $f(\tau)$ where $h(-\tau)$ is the weighting function [36].

By the Fourier theory, a given signal can be synthesized as a summation of sinusoidal waves of various amplitudes, frequencies and phases [35]. In other words, a time domain signal is represented by an amplitude spectrum and a phase spectrum using the Fourier transform in the frequency domain [35]. Convolution in the time domain is equivalent to a point-wise multiplication of the amplitude spectra and an addition of the phase spectra in the frequency domain [36]. Thus, noise, if convolved with the signal, can be more easily separated from the signal in the frequency domain. In signal and image processing, a filter is commonly developed to filter out such noises based on their frequency differences as compared with the signal, and the inverse Fourier transform is applied afterwards to transform the true signal back to the time domain. Deconvolution is a filtering process which removes a noisy wavelet from the recorded data by reversing the process of convolution [36].

In summary, Fourier transform and its inverse form a pair of mathematical operators that transform signals across the time domain and the frequency domain. Convolution and deconvolution form a pair of reversal processes, i.e., generating a third signal by superimposing two signals on top of each other via convolution, and decomposing one signal into two through deconvolution. Convolution and deconvolution are operators named in the time domain, representing the processes occurred in the frequency domain.

Deconvolution methods

In image processing, a point source of light is considered distorted by convolving with the PSF of the imaging system [37], as illustrated in Figure 2.4. In economics, the reaction of economy is considered convolved with the external or internal shocks. The assumption here is that PSF is isoplanatic, i.e., the noise is symmetrically distributed in the term of economics. The inefficiency, on the other hand, is assumed to be positive, as the pixels in an image can not be negative. These allow the deconvolution system being a perfect model for inefficiency decomposition, which is applied in this thesis to identify the PSF or shocks from the true image (image processing) or behavior (economics). Considering a dynamic stochastic system, these problems can be mathematically expressed as (2.1), where $f(x)$ is the true signal, $h(x)$ is PSF and $\varepsilon(x)$ is the random noise.

$$g(x) = f(x) \otimes h(x) + \varepsilon(x) \tag{2.1}$$

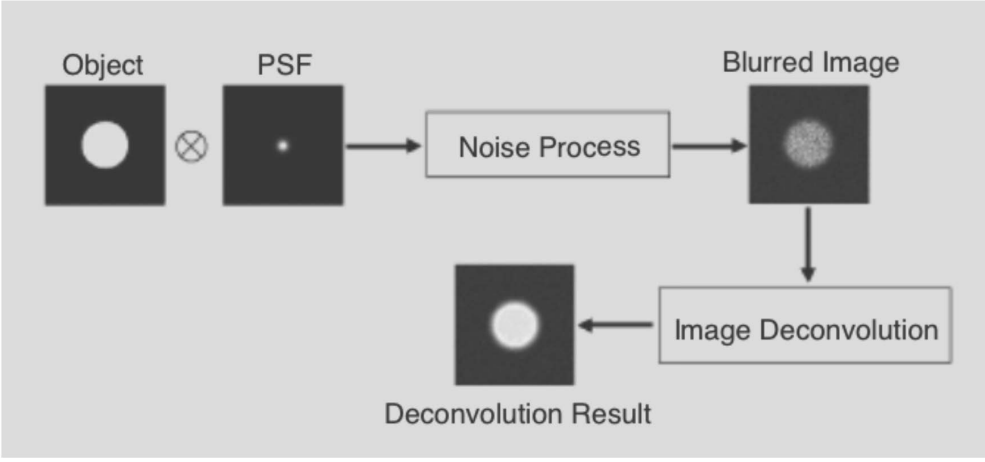


Figure 2.4. Schematic of a general deconvolution procedure.

Such problems are easy to solve (by directly applying deconvolution to the convolved data) if PSF is known. This refers to classical deconvolution which comprises of a large body of techniques and has matured since its inception in the 1960s [38, 39]. These approaches differ primarily in the prior information they include to perform the restoration task. The earliest algorithms to tackle the blind deconvolution problem appeared in mid 1970s [40, 41], which attempted to identify known patterns in

the blur. A small but dedicated effort followed in the late 1980s [42–46] and a resurgence was seen in the 1990s [47, 48].

In practice, it is often costly, dangerous or physically impossible to obtain *a priori* information on the true object or PSF. For example, in astronomy, it is difficult to model the original image which has not been imaged before; in addition, the degradation from blurring can not be properly specified [49]. Since 1990s, the area has been extensively explored with many blind deconvolution methods developed. Blind deconvolution, as stated by its name, is the process of recovering the blurred image in the presence of a poorly determined or unknown PSF. Many algorithms have been developed accordingly, which can be roughly classified into five categories. These methods, differing mainly by the assumptions they made on $f(x)$ and $h(x)$, are 1) *a priori* blur identification methods [41], 2) zero sheet separation methods [50, 51], 3) autoregressive moving average (ARMA) parameter estimation methods [52], 4) nonparametric methods based on high order statistics (HOS) [53, 54], and 5) nonparametric iterative methods [44, 55, 56].

The *a priori* methods identify the PSF before performing blind deconvolution, which typically assume a known parametric form for the PSF. This class of deconvolution methods is relatively simple to implement and computationally less complex as compared with other approaches. However, it requires the prior knowledge of the form of PSF and is sensitive to the additive noise term.

In zero sheet separation methods, the analytically continued Fourier transform of a two-dimensional image vanishes to zero on a two-dimensional surface, which uniquely characterizes the image and is called a zero sheet. Instead of manipulating a function in multiple-dimensional space, the projections of zero sheets were calculated (named zero tracks) and used for retrieving the true image and PSF. Techniques as such outweigh the other methods by providing valuable insights into the blind deconvolution problems in multiple dimensions. However, they are highly sensitive to noise and prone to inaccuracy for larger images since the noise term is dropped in the Z-transform of (2.13), as shown in (2.2).

$$G(\omega) = F(\omega) \cdot H(\omega) \tag{2.2}$$

ARMA parameter estimation methods model the blurred image as a ARMA process, i.e., modeling the true image as an autoregressive (AR) process and the PSF as a moving average (MA) process. The advantage associated with these approaches is the noise-insensitive nature since the noise is already taken into account by the model. However, these methods have the risk of ill-convergence, and the total number of parameters can not be very large for practical computations. Also, the deconvolution results are often not unique unless additional assumptions are made on the PSFs.

Nonparametric methods based on HOS minimize the given cost function that accounts for the probabilistic non-Gaussian nature of the true image. Specifically, the recorded image is passed through a finite

impulse response (FIR) inverse filter, which yields an estimate of the true image. The parameters of the FIR filter are updated accordingly to optimize the function that incorporates the HOS model of the true image. Algorithms belonging to this class include, e.g., minimum entropy deconvolution (MED) [57] and Bayesian non-minimum phase approaches [54]. The primary advantages of these methods are that they can identify non-minimum phase PSFs and are robust to noise. However, these approaches require accurate modeling of the true image by a known non-Gaussian probability distribution and the algorithms may be trapped in local minima in the estimation process.

Nonparametric iterative methods do not require certain parametric form for the true image or the PSF. Algorithms belonging to this category include, e.g., iterative blind deconvolution (IBD) [44], simulated annealing (SA) [55], and non-negativity and support constraints recursive inverse filtering (NAS-RIF) [56]. There are two common features for approaches of this kind. First, they generally assume certain constraints on the original image and PSF. Typical constraints in the spatial domain are 1) the true image is non-negative, 2) the background image is uniformly black, gray or white, 3) the support size of the original object is known. Second, they all employ iterative methods to minimize a cost function with respect to the forward or inverse filter coefficients. Different nonparametric iterative approaches differ in the objective of minimization and how the cost function is constructed. For example, the cost functions of IBD and SA are minimized with respect to both $f(x)$ and $h(x)$ simultaneously, and that of NAS-RIF optimizes the coefficients of the inverse filter $h^{-1}(x)$ that convolves with the blurred image to estimate the original image (2.3). The main advantage of these methods is that they do not require any prior knowledge on either the original image or PSF except for the support size. However, the cost function is not necessarily convex and thus may not always guarantee a global optimization. In particular, IBD and SA are relatively robust to noise, and NAS-RIF is guaranteed to achieve the global minimal.

$$\hat{f}(x) = g(x) \otimes h^{-1}(x). \tag{2.3}$$

The definitions of the 5 categories of blind deconvolution methods are summarized in Table 2.1, and their characteristics are compared in Table 2.2 [47].

Richard-Lucy blind deconvolution

Richard-Lucy blind deconvolution (RLb), applied in this thesis for inefficiency estimation and well described in (Publication II), belongs to the category of IBD and implies all the assumptions held by IBD (see the afore section for details). Its iterative nature requires us to set the convergence criteria to stop the algorithm, which could be the minimum change on the parameters or simply the maximum number on the iterations. In this thesis, the convergence was set to 10 rounds of iterations based on its

Table 2.1. Definitions of major blind deconvolution methods.

Class of methods	Definition
a priori blur identification methods	Algorithms that estimate the PSF priori to image restoration using known characteristics of the PSF and true image.
Zero sheet separation methods	Algorithms that perform blind deconvolution by factoring the two-dimensional Zero sheet of the blurred image.
ARMA parameter estimation methods	Algorithms that model the blurred image using an ARMA model and perform deconvolution by estimating these parameters.
Nonparametric methods based on HOS	Algorithms that make use of high order statistics information about the image for restoration.
Nonparametric iterative methods	Algorithms that make deterministic assumptions on the image and PSF, and estimate them by iteratively minimising a cost function with respect to the forward or inverse filter coefficients.

convention.

In this thesis, the performance of the RLb method is compared with the MM method. The MM method computes the inefficiencies according to the Jondrow's method (section 3.3.3 of [22]), where the parameters are computed by the method of moment (section 3.3.1 of [22]). The simulation results of RLb as compared with MM are presented in Figures 2 and 3 of (Publication II), where the true and estimated inefficiencies obtained from RLb (red circle) and MM (blue star) are plotted along the x and y axes, respectively. 'NA's, missing outputs from MM due to the wrong skewness assumption, are omitted from the plot. From these results, it is clear that RLb outperforms MM in its robustness to 1) distribution skewness and kurtosis, 2) distribution assumption, 3) data noise and 4) heteroscedasticity as described in (Publication II). In certain cases MM shows superiority over RLb (Figure 3 (d) and (e)), however, we should note that 90% of the MM estimates are non-valid even with the modest inefficiency levels (Table 3); also, the large STD regarding DMUs' deviation from the true inefficiencies may change their rank, leading to poor performance in benchmarking. It is also worthwhile to address the issue of shrinkage here. [58] and [59] have pointed out that we will overestimate u_i when it is small, and underestimate it when u is large and of half or truncated normal distribution. Here, we indeed observed shrinkage on $\hat{u}_i$ estimated using MM when u_i is of truncated normal (Figure 3 (c)), and a slight trend towards shrinkage when u_i is of half normal distribution (Figure 3 (a)). This might because that all the simulated data points in Figure 3 (a) are below 1, which could not be considered large and thus be overestimated using the MM approach overall. Also note that, we do not observe shrinkage in homoscedastic scenarios where all data points are generated from one distribution and should not vary regarding the bias towards the true value. In RLb, the parameters are updated iteratively, i.e., u_i is not simply estimated from ε and does not satisfy the precondition of the shrinkage issue. The RLb method, once adjusted for the panel data, is named cRLb in this thesis. The cRLb estimates the inefficiency at each time point for each firm, which fits both time-varying and time-invariant panel settings. Also worth mentioning is that, though we show

Table 2.2. Characteristics of major blind deconvolution methods.

Class of methods	*a priori* blur identification methods	Zero sheet separation methods	ARMA parameter estimation methods	Nonparametric methods based on HOS	Nonparametric iterative methods
Assumptions on true image	possibly contains edges or point sources	finite support	modelled by an AR process	accurately modelled by a non-Gaussian probability distribution	deterministic constraints such as non-negativity, support, blur invariant edges
Assumptions on PSF	symmetric and non-minimum phase with a possibly known parametric form	finite support	symmetric and modelled by an MA process of a possibly known parametric form	invertible	IBD and SA (positive with known finite support), NAS-RIF (invertible)
Complexity	Very low	High	moderate to high	moderate	low to moderate
Convergence	not iterative	sensitive to numerical inaccuracies, results in ill-convergence	ill-convergence to local minima, sensitive to initial conditions	ill-convergence, sensitive to initial estimate	IBD (ill-convergence, sensitive to initial estimate), SA and NAS-RIF (converge to global minima)
Sensitivity to noise	moderate to high	high	moderate	low (Gaussion)	IBD (low), SA (moderate), NAS-RIF (moderate to high)
Conventional application area	astronomy, industrial x-ray imaging, photography	astronomy	photography, texture image reconstruction	astronomy, seismic data analysis	magnetic resonance imaging, position emission tomography, x-ray imaging, astronomy

the advantages of RLb and cRLb as compared with the traditional inefficiency estimation approaches, we could not make any statistical statement on its accuracy without formal mathematical proof which is left for the future studies.

2.2 Benchmarking and clustering

2.2.1 Benchmarking

Benchmarking is the process comparing the performance or activities of one unit against that of the 'best practice' units. Productive efficiency is a natural parameter for such performance assessment. Many existing statistical methods can be used for productive efficiency analysis including both parametric and non-parametric approaches. CNLS, corrected CNLS (C^2NLS), DEA, and StoNED are non-parametric methods; ordinary least squares (OLS), corrected OLS (COLS), parametric programming (PP), and SFA are parametric ones. Among these techniques, OLS, COLS, PP, C^2NLS and DEA are static which attribute all deviations to the inefficiencies; while SFA and StoNED are stochastic which take noise into account [22]. Out of these approaches, DEA has gained its popularity in benchmarking due to its non-parametric nature and flexibility. That is to say, DEA does not require a specific functional form of the relationship between the inputs and outputs, and it allows for multiple dimensional inputs and different combinations of products and services to be equally attractive [60]. In addition, DEA directly outputs one or a few benchmarks for each decision making unit (DMU) besides the efficiency scores [60]. Although traditional DEA has been widely applied in benchmarking, it does not consider the circumstance under which each DMU operates, rendering the benchmarks obtained quite often impossible to achieve [61]. This leads to the advent of context-dependent DEA which peels off the frontier in a sequential fashion to cluster DMUs into groups at different efficiency levels [62]. This method [62] and its variants [63, 64] allow DMUs finding their achievable goals at each efficiency stage, but do not take into account the segment heterogeneity. This makes the benchmarking of a DMU from one segment to another most often impractical, as DMUs belonging to different segments may significantly differ in, e.g., the operational structure. Thus, a segment-specific benchmarking strategy is called for, which not only outputs the targets achievable in a step-wise manner but also realistic in the long run.

2.2.2 Clustering

Clustering is the task of grouping a set of objects in a way that objects in the same group (namely a cluster) are more similar to each other than to those in other groups. It is a main task of data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics. Clustering is used in this thesis to classify DMUs into properly defined groups so that the benchmarks are identified (according to the efficiencies obtained from productive efficiency analysis) among DMUs with similar operational structure. Below, some basic concepts and traditional methods used in clustering are introduced.

Clustering concepts

Clustering is the most important unsupervised machine learning method, which identifies a structure from a collection of unlabeled data. A cluster could be defined as a collection of objects which are similar among themselves but dissimilar to the objects belonging to the other clusters. Such a concept is explained in Figure 2.5

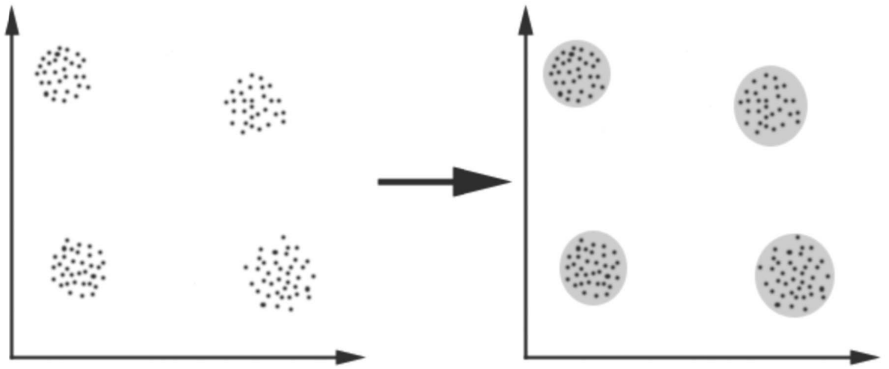


Figure 2.5. Illustration of the concept of clustering.

Clustering itself does not refer to a particular algorithm, but a general task, which can be achieved by various algorithms that differ significantly in their concept of what constitutes a cluster and how to efficiently find them. Popular concepts of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions, etc. There is no absolute 'best' criterion on the number of clusters and where the group boundaries should be drawn. The choice of clustering always depends on its final aim, i.e., the user must supply this criterion in such a way that the clustering result suits his/her needs. For instance, one might be interested in identifying representatives for homogeneous groups (dimension reduction), finding 'natural clusters' that describe their distributional properties (data type identification), unveiling useful groupings (data class subtyping) or identifying unusual data objects (outlier detection).

Clustering methods

Many methods have been developed for clustering [65–67], with the most commonly used approaches roughly classified into three categories, i.e., hierarchical methods, partitioning methods, and model-based methods [68].

Hierarchical methods can be either agglomerative or divisive, which proceeds by recursively fusing or separating the objects into greater or finer groups to optimize a certain criterion [65]. Different criteria are developed to serve this purpose, among which single linkage [69], complete linkage [69], average linkage [69], group average linkage [69], and Ward's linkage [70] are widely applied [69] (formulas are shown in (2.4) to (2.8). Distances such as Euclidean distance [71], Mahalanobis distance [72], Manhattan distance [73], and Hamming distance [74] are generally adopted in these criteria to measure the cluster dissimilarity. These distances can be computed from (2.9) to (2.12), respectively, where p ($p \in \{1, \infty\}$) is the dimension of each observation and 'Cov' represents the covariance matrix of two objects (firms are represented as objects here). The accuracy of hierarchical clustering highly depends on the distance measurement, which requires expert domain knowledge especially for complex data types. For example, Euclidean distance, which is commonly used when data is representable in vector space, is not appropriate for high-dimensional text clustering [75]; and semantic similarity measurements, such as graph-structure based distances and information content based methods, are especially applicable to gene ontology (GO) based clustering [76]. Further, hierarchical clustering is computationally inefficient, given that computing distances among all observation pairs requires a complexity of $O(n^2)$, where n is the number of observations [77]. Also, at what granularity should the algorithm stop is an important issue and could not be naturally determined without prior knowledge or estimation on the number of

clusters [68].

$$D(G_i, G_j) = \min_{\mathbf{r}_a \in G_i, \mathbf{r}_b \in G_j} d(\mathbf{r}_a, \mathbf{r}_b) \tag{2.4}$$

$$D(G_i, G_j) = \max_{\mathbf{r}_a \in G_i, \mathbf{r}_b \in G_j} d(\mathbf{r}_a, \mathbf{r}_b) \tag{2.5}$$

$$D(G_i, G_j) = \frac{\sum_{a=1}^{N_{G_i}} \sum_{b=1}^{N_{G_j}} d(\mathbf{r}_a, \mathbf{r}_b)}{N_{G_i} \times N_{G_j}} \tag{2.6}$$

$$D(G_i, G_j) = d(\frac{\sum_{a=1}^{N_{G_i}} \mathbf{r}_a}{N_{G_i}}, \frac{\sum_{b=1}^{N_{G_j}} \mathbf{r}_b}{N_{G_j}}) \tag{2.7}$$

$$\begin{aligned} D(G_i, G_j) &= ESS(G_iG_j) - ESS(G_i) - ESS(G_j), \text{ where} \\ ESS(G_i) &= \sum_{a=1}^{N_{G_i}} |\mathbf{r}_a - \frac{1}{N_{G_i}} \sum_{w=1}^{N_{G_i}} \mathbf{r}_w|^2 \end{aligned} \tag{2.8}$$

$$d(\mathbf{r}_a, \mathbf{r}_b) = \sqrt{\sum_{w=1}^{p} (r_{aw} - r_{bw})^2} \tag{2.9}$$

$$d(\mathbf{r}_a, \mathbf{r}_b) = \sqrt{(\mathbf{r}_a - \mathbf{r}_b)^T \mathrm{Cov}^{-1} (\mathbf{r}_a - \mathbf{r}_b)} \tag{2.10}$$

$$d(\mathbf{r}_a, \mathbf{r}_b) = \sum_{w=1}^{p} |r_{aw} - r_{bw}| \tag{2.11}$$

$$d(\mathbf{r}_a, \mathbf{r}_b) = \sum_{w=1}^{p} \kappa_w, \quad \kappa_w = \begin{cases} 1 & \text{if } r_{aw} \neq r_{aw} \\ 0 & \text{if } r_{aw} = r_{bw} \end{cases} \tag{2.12}$$

Partitioning methods belong to another class of heuristic methods besides hierarchical clustering. The principle is to iteratively reallocate data points across groups until no further improvement is obtainable [66, 68]. K-means [66] is a typical and the most representative partitioning algorithm. It is based on the criterion that each object belongs to its closest group, where the group is represented by the mean of its objects. In particular, with a given g, the algorithm partitions N observations, $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$, into g groups ($\mathbf{G} = \{G_1, G_2, \ldots G_g\}$) by minimizing the total intra-cluster variance, i.e., $\underset{\mathbf{G}}{\operatorname{argmin}} \sum_{i=1}^{g} \sum_{\mathbf{r}_w \in \mathbf{G}_i} (\mathbf{r}_w - \mu_i)^2$, where μ_i is the average of $\mathbf{G}_i$.

It is seen from K-means that the number of clusters has to be pre-specified or known. Also, the clustering results may be contaminated by outliers [68]. Successive efforts have been devoted to search their

remedies which, however, mostly involve techniques out of the domain of partitioning methods. For example, X-means (extended from K-means) solves the problem of selecting the number of clusters via using model selection criteria [78].

Despite those disadvantages, partitioning methods are widely applied due to their simplicities. Many algorithms, such as fuzzy C-means [79], quality threshold clustering [80] and partitioning around medoids [81], also belong to this category. Specifically, 'fuzzy C-means' assigns each data point to each cluster with a certain probability [79]; 'quality threshold' only groups data points whose similarities are high enough [80]; and 'partitioning around medoids' minimizes a sum of dissimilarities and allows the user to choose the number of clusters through graphical display [81].

Hierarchical methods and partitioning methods are also called 'heuristic methods', both of which rely on some heuristics and follow intuitively reasonable procedures [68]. Although considerable research has been done on these methods, still little associated systematic guidance is available for solving some practical issues [68]. These include how to specify the number of clusters, how to handle the outliers, and how to choose or define a good distance for a particular clustering problem.

Model based methods attempt to optimize the fitness between the data and the model where the data is assumed to be generated [67, 77, 82, 83]. Model based methods can be further classified into finer groups, including finite mixture models [67], infinite mixture models [82], model based hierarchical clustering [83], and specialized model based partitioning clustering [77] (e.g., Self Organizing Map (SOM) [84]), among which finite model based methods are most widely applied.

In finite model based clustering, each observation $\mathbf{r}$ is drawn from a finite mixture distributions with the prior probability π_i, component-specific distribution f_i and its parameters θ_i. The formula is given by

$$f(r;\Theta) = \sum_{i=1}^{g} \pi_i f_i(r;\theta_i), \tag{2.13}$$

where $\Theta = \{(\pi_i, \theta_i) : i = 1, \ldots, g\}$ is used to denote all unknown parameters, and $0 \leq \pi_i \leq 1$ for any i and $\sum_{i=1}^{g} \pi_i = 1$. Note that g is the number of components in this model.

Expectation maximization (EM) algorithm is normally used for the above model-based clustering. The data log-likelihood can be written as

$$\log L(\Theta) = \sum_{j=1}^{N} \log(\left[\sum_{i=1}^{g} \pi_i f_i(r_j;\theta_i)\right]), \tag{2.14}$$

where $R = \{r_j : j = 1, ..., n\}$ and N is the total number of observations.

Since direct maximization of (2.14) is difficult, the problem can be casted in the framework of incomplete data. Define I_{ji} as the indicator of whether r_j comes from component i, the complete data log-likelihood

becomes

$$\log L_c(\Theta) = \sum_{j=1}^{N} \sum_{i=1}^{g} I_{ji} \log (\pi_i f_i(r_j; \theta_i)). \tag{2.15}$$

At the m^{th} iteration of the EM algorithm, the E step computes the expectation of the complete data log-likelihood which is denoted as Q

$$\begin{aligned} Q(\Theta; \Theta^{(m)}) &= E_{\Theta^{(m)}}(\log L_c | R) \\ &= \sum_{j=1}^{N} \sum_{i=1}^{g} \tau_{ji}^{(m)} \log (\pi_i f_i(I_j; \theta_i)), \end{aligned} \tag{2.16}$$

and the M step updates the parameter estimates to maximize Q. The algorithm is iterated until convergence. Note that I's in (2.15) are replaced with τ's in (2.16), and the relationship between these two parameters is $\tau_{ji} = E[I_{ji}|r_j, \hat{\theta}_1, ..., \hat{\theta}_g; \hat{\pi}_1, ..., \hat{\pi}_g]$. The set of parameter estimates $\left\{\hat{\theta}_1, ..., \hat{\theta}_g; \hat{\pi}_1, ..., \hat{\pi}_g\right\}$ is a maximizer of the expected log-likelihood for given τ_{ji}'s, and we can assign each r_j to its component based on $\left\{i_0 | \tau_{ji_0} = \max_i \tau_{ji}\right\}$.

One advantage of mixture model based clustering is its automatic determination of the number of clusters. Commonly used model selection criteria can be roughly classified as likelihood-based methods [85] and approximation-based methods [86–91]. Four approximation-based model selection criteria are widely applied due to their computational efficiency, which are Akaike information criterion (AIC) [87, 90], modified AIC (AIC3) [89, 90], Bayesian information criterion (BIC) [88, 91], and integrated classification likelihood-BIC (ICL-BIC) [86].

Model based methods can naturally solve the problems generically inherited by heuristic methods [68] which, e.g., often determine the number of clusters by casting it as the model selection problem and group the outliers as separate clusters [67, 68]. Further, model based methods outweigh heuristic methods in their statistical nature [67, 68].

Finite model based clustering is used in this thesis given its statistical nature. In finite mixture model gene clustering, each observation $\mathbf{x}_j$ ($j = 1, \ldots, n$ and n is the number of genes) is assumed to be drawn from finite mixture distributions with the prior probability π_δ, component-specific distribution f_δ and its parameters θ_δ [67]. The formula is shown in (2.17) [67], where $\theta = \{(\pi_\delta, \theta_\delta) : \delta = 1, \ldots, g\}$ represents all the unknown parameters, $0 < \pi_\delta \leq 1$ for any δ, and $\sum_{\delta=1}^{g} \pi_\delta = 1$.

$$f(\mathbf{x}_j|\theta) = \sum_{\delta=1}^{g} \pi_\delta f_\delta(\mathbf{x}_j|\theta_\delta) \tag{2.17}$$

NMM-StoNED

In this thesis, NMM-StoNED is proposed for clustering-based benchmarking. Two approaches are proposed, i.e., 1) frontier estimation followed by clustering (estimate-cluster-benchmark) and 2) clustering followed by frontier estimation (cluster-estimate-benchmark). These two approaches differ in their assumptions. The 'estimate-cluster-benchmark' approach assumes that all firms have access to and use the same production technology, and benchmarking is performed over these identified peers. The 'estimate-benchmark-cluster' approach, on the other hand, assumes that the differences across firms are at the technology level which is taken into account at the frontier estimation stage. In addition, the second approach needs comparatively more DMUs to apply, with sufficient DMUs fell in each cluster for frontier estimation. Here 'sufficient' means that the number of DMUs meets the minimum requirement of the frontier estimation method such as StoNED. The users should decide on which approach to use according to the underlying assumptions of each problem. If the assumptions required for the 'estimate-cluster-benchmark' framework hold, but the user mis-applied the 'cluster-estimate-benchmark' framework, then the estimates for the individual groups are inefficient, whereas the proposed estimator is biased the other way around. No matter which approach the user chooses, the variables for clustering should be pre-selected in the clustering stage. This needs our prior knowledge on, e.g., the operational structure of firms and traditional efficiency measures.

The strategy of including a clustering stage in benchmarking is close to latent class stochastic frontier analysis [92]. However, these two methods differ in the assumptions where the heterogeneity comes from. Specifically, latent class SFA models the heterogeneity using a latent class structure, which assumes that the differences come from the inefficiencies and noise; while the clustering stage of our proposed framework uses an unsupervised technique to model the differences according to the measures it takes as the inputs.

The choice of combining NMM and StoNED in this thesis for clustering-based benchmarking is rested on the following bases. StoNED is selected as the frontier estimator because it is considered the best practice for benchmark regulation of electricity distribution [24], which is suitable for our empirical setup. NMM is chosen as the clustering approach due to its superiority over, e.g., the most popular clustering technique K-means, when combined with StoNED in our study (Publication IV). It would be interesting to try other combinations in this framework under different problem settings.

3. Summary

This thesis focuses on interdisciplinary application of computational methods in quantitative economics, with efforts devoted to two topics, i.e., 'decompose inefficiency from composite errors using deconvolution methods' and 'find segment-specific benchmarks using clustering techniques' (Table 3.1).

Table 3.1. Summary of the thesis.

Techniques	Conventional use	Problem	Article	Content
Deconvolution	Noise decomposition in image processing	Efficiency estimation	Publication I	Framework
			Publication II	Combination choice
Clustering	Classification in gene grouping	Benchmarking	Publication III	Cross-sectional case
			Publication IV	Panel case

In the first topic, the Richardson-Lucy blind deconvolution (RLb) method is used to decompose inefficiencies from the composite errors in the cross-sectional setting (Publication I) and the corrected RLb (cRLb) is proposed to solve such problems in the panel setting (Publication II). The RLb method outweighs conventional methods such as MM in at least five aspects. First, it is non-parametric. Second, it never outputs null or zero values due to incorrect skewness or low kurtosis of inefficiency density. Third, it is insensitive to the distributional assumption of the inefficiency term u. Fourth, it is robust to data noise level. Fifth, it is insensitive to data heteroscedasticity. The cRLb method inherits all the merits of RLb, and estimates the inefficiency for each DMU at each time point.

In the second topic, clustering-based benchmarking framework (Publication III), particularly NMM-StoNED (Publication IV), is proposed to take into account the heterogeneity of firms and their operating environment in benchmarking. This framework novels in the following four aspects. First, it adjusts benchmarking according to the intrinsic characteristics of DMUs. Second, it is highly flexible in a sense that 'clustering' and 'efficiency estimation' can be tuned or optimized, separately. The efficiencies can be computed using different frontier models and the inputs can be customized depending on the factors users wish to evaluate. Also, the algorithms at each step could be freely chosen, modified or developed to meet the customer needs, allowing more freedom to the users and a better chance of getting the optimal targets. Third, it provides multiple absolute benchmarks for the inefficient DMUs to choose, and ensures at least one relative benchmark for each DMU in cases where no DMU achieves 100% efficiency. By comparing different combinations of clustering techniques and efficiency estimation methods, NMM-StoNED is proposed given its superior performance as evaluated using both Monte Carlo simulations and empirical study.

4. Conclusion

Interdisciplinary application of methods across fields casts novel view on the problems which are hardly approachable using conventional methods and, quite often, brings surprisingly good solutions. Efforts encompassed in this thesis show at least the following two advantages of interdisciplinary application of computational methods in quantitative economics.

- It solves the problems hardly approachable using conventional methods, such as issues derived from the dependence on the distributional assumption in inefficiency analysis (Publication I and Publication II).

- It improves the current techniques in addressing problems such as the heterogeneities among firms in benchmarking (Publication III and Publication IV).

The developed interdisciplinary methods are shown to be efficient tools in solving the afore-discussed problems in quantitative economics. It would be interesting and useful to deploy these developed methods to tackle more empirical problems which may offer high values in practice. Also, with the successful stories demonstrated in this thesis, it is worthwhile to explore more techniques commonly applied in other fields to improve current methodologies in quantitative economics and solve the corresponding problems.

Bibliography

[1] Mas-Colell A., Whinston MD, Green JR (1995) Microeconomic Theory. Oxford University Press, New York, USA.

[2] Syverson C (2011) What determines productivity? Journal of Economic Literature, 49 (2): 326-365.

[3] Kuosmanen T and Fosgerau M (2009) Neoclassical versus frontier production models? testing for the skewness of regression residuals. Scandinavian Journal of Economics, 111 (2): 351-367.

[4] Abramowitz M (1956) Resource and output trends in the United States since 1870. American Economic Review, 46 (2): 5-23.

[5] Dyson RG, Allen R, Camacho AS, Podinovksi VV, Sarrico CS, Shale EA (2001) Pitfalls and protocols in DEA. European Journal of Operational Research, 132: 245-259.

[6] Farrell MJ (1957) The measurement of productive efficiency. Journal of the Royal Statistical Society: Series A, 120 (3): 253-282.

[7] Charnes A, Cooper WW, Rhodes E (1978) Measuring the inefficiency of decision making units. European Journal of Operational Research, 2 (6): 429-444.

[8] Aigner DJ, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier models. Journal of Econometrics, 6: 21-37.

[9] Meeusen W, van den Broeck J (1977) Efficiency estimation from cobb-Douglas production function with composed error. International Economic Review, 8: 435-444.

[10] Sauer J (2006) Economic theory and econometric practice: parametric efficiency analysis. Empirical Economics, 31: 1061-1087.

[11] Fried H, Lovell CAK, Schmidt S (2008) The measurement of productive efficiency and productivity change. Oxford University Press, New York.

[12] Gong B-H, Sickles R C (1992) Finite sample evidence on the performance of stochastic frontiers and data envelopment analysis using panel data. Journal of Econometrics, 51: 259-284.

[13] Banker RD, Gadh VM, Gorr WL (1993) A Monte Carlo comparison of two production frontier estimation methods: corrected ordinary least squares and data envelopment analysis. European Journal of Operational Research, 67: 332-343.

[14] Andor M, Hesse F (2013) The StoNED age: the departure into a new era of efficiency analysis? A Monte Carlo comparison of StoNED and the 'oldies' (SFA and DEA). Journal of Productivity Analysis, DOI 10.1007/s11123-013-0354-y.

[15] Fan Y, Li Q, Weersink A (1996) Semiparametric estimation of stochastic production frontier models. Journal of Business & Economic Statistics, 14 (4): 460-468.

[16] Kneip A, Simar L (1996) A general framework for frontier estimation with panel data. Journal Productivity Analysis, 7: 187-212.

[17] Kumbhakar SC, Park BU, Simar L, Tsionas EG (2007) Nonparametric stochastic frontiers: a local maximum likelihood approach. Journal of Econometrics, 137: 1-27.

[18] Simar L, Zelenyuk V (2008) Stochastic FDH/DEA estimators for frontier analysis. Journal of Productivity Analysis, doi:10.1007/s11123-010-0170-6.

[19] Banker RD, Maindiratta A (1992) Maximum likelihood estimation of monotone and concave production frontiers. Journal of Productivity Analysis, 3: 401-415.

[20] Kuosmanen T (2008) Representation theorem for convex nonparametric least squares. Economic Journal, 11: 308-325.

[21] Kuosmanen T, Johnson A (2010) Data envelopment analysis as nonparametric least squares regression. Operational Research, 58 (1): 149-160.

[22] Kuosmanen T, Kortelainen M (2010) Stochastic non-smooth envelopment of data: semiparametric frontier estimation subject to shape constraints. Journal of Productivity Analysis, do:10.1007/s11123-010-0201-3.

[23] Schmidt P, Sickles R (1984) Production frontiers and panel data. Journal of Business & Economic Statistics, 2: 367-374.

[24] Kuosmanen T, Saastamoinen A, Sipilälinen T (2013) What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods. Energy Policy, 61: 740-750.

[25] Porembski M, Breitenstein K, Alpar P (2005) Visualizing efficiency and reference relations in data envelopment analysis with an application to the branches of a German bank, Journal of Productivity Analysis, 23 (2), 203-221.

[26] Silva Portela MCA, Thanassoulis E (2005) Profitability of a sample of Portuguese bank branches and its decomposition into technical and allocative components. European Journal of Operational Research, 162 (3), 850-866.

[27] Davis S, Albright T (2004) An investigation of the effect of balanced scorecard implementation on financial performance. Management Accounting Research, 15 (2): 135-153.

[28] Camanho AS, Dyson RG (2005) Cost efficiency, production and value-added models in the analysis of bank branch performance, Journal of the Operational Research Society, 56 (5): 483-494.

[29] Agrell P, Bogetoft P, Tind J (2005) DEA and dynamic yardstick competition in Scandinavian electricity distribution. Journal of Productivity Analysis, 23 (2): 173-201.

[30] Delmas M, Tokat Y (2005) Deregulation, governance structures, and efficiency: the U.S. electric utility sector, Strategic Management Journal, 26 (5): 441-460.

[31] Pollitt M (2005) The role of efficiency estimates in regulatory price reviews: Ofgem's approach to benchmarking electricity networks, Utilities Policy, 13 (4): 279-288.

[32] Haykin S (1993) Blind deconvolution, S S Haykin and T Kailath, Prentice Hall, New York.

[33] Hamilton J (1994) Chapter 1. In 'Time Series Analysis', Princeton University Press.

[34] Lütkepohl H (2008) Impulse response function. In 'The New Palgrave Dictionary of Economics', S N Durlauf and L E Blume, Palgrave Macmillan.

[35] Bracewell R (1986) The Fourier transform and its applications, McGraw-Hill Science/Engineerin/Math.

[36] Sobolev VI (2001) Convolution of functions. In 'Encyclopedia of Mathematics', Hazewinkel M, Springer.

[37] Cheng PC (2006) The contrast formation in optical microscopy. In 'Handbook of Biological Confocal Microscopy', J B Pawley, Springer.

[38] Katsaggelos A K (1991) Digital Image Restoration. Springer-Verlag.

[39] Banham MR, Katsaggelos AK (1997) Digital image restoration. IEEE Signal Processing Magagine, 14 (2): 24-41.

[40] Stockham TG, Cannon TM, Ingebretsen RB (1975) Blind deconvolution through digital signal processing. Proceedings IEEE, 63 (4): 678-692.

[41] Cannon M (1976) Blind deconvolution of spatially invariant image blurs with phase. IEEE Transactions on Acoustics Speech and Signal Processing, 24 (1): 58-63.

[42] Tekalp AM, Kaufman H, Woods JW (1986) Identification of image and blur parameters for the restoration of noncausal blurs. IEEE Transactions on Acoustic, Speech, and Signal Processing, 34: 963-972.

[43] Lane RG, Bates RHT (1987) Automatic multidimensional deconvolution, Journal of the Optical Society of America-A, 4 (1): 180-188.

[44] Ayers G, Dainty J (1988) Iterative blind deconvolution method and its applications. Optics Letters, 13 (7): 547-549.

[45] Lay KT, Katsaggelos AK (1988) Simultaneous identification and restoration of images using maximum likelihood estimation and the EM algorithm, Proceeding of 26th Annual Allerton Conference on Communication, Control and Computing (Monticello, IL), 661-662.

[46] Lagendijk RL, Biemond J, Boekee DE (1989) Blur identification using the expectation-maximization algorithm, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 3: 1397-1400.

[47] Kundur D, Hatzinakos D (1996) Blind image deconvolution. IEEE Signal Processing Letters, 13 (3): 43-64.

[48] Kundur D, Hatzinakos D (1996) Blind image deconvolution revisited. IEEE Signal Processing Magazine, 13 (6): 61-63.

[49] Pantin E, Starck J, Murtagh F (2007) Deconvolution and blind deconvolution in astronomy. In 'Blind Image Deconvolution: Theory and Applications', P Campisi and K Eqiazarian, CRC Press.

[50] Ghiglia D, Romero L, Mastin G (1993) Systematic approach to two-dimensional blind deconvolution by zero-sheet separation. Journal of the Optical Society of America, 10 (5): 1024-1036.

[51] Premaratne P (1999) Zero sheet separation of blurred images with symmetrical point spread functions. Conference Record of the Thirty-third Asilomar Conference on Signals, Systems & Computers, 2:i-xxiv.

[52] Lagendijk R (1990) Maximum likelihood image and blur identification: a unifying approach, Optical Engineering, 52 (3): 345-370.

[53] Wu HS (1990) Minimum entropy deconvolution for restoration of blurred two-tone images. Electronics Letters, 26 (15): 1183-1184.

[54] Jacovitti G, Neri A (1990) A Bayesian approach to 2D non minimum phase AR identification. Fifth ASSP Workshop on Spectrum Estimation and Modeling, 79-83.

[55] McCallum B (1990) Blind deconvolution by simulated annealing. Optics Communications, 75 (2): 101-105.

[56] Kundur D, Hatzinakos D (1998) Novel blind deconvolution scheme for image restoration using recursive filtering. IEEE Transactions on Signal Processing, 46 (2): 375-390.

[57] Wiggins RA (1978) Minimum entropy deconvolution. Geoexploration, 16 (1-2): 21-35.

[58] Wang WS, Schmidt P (2002) On the distribution of estimated technical efficiency in stochastic frontier models. Journal of Productivity Analysis, 148(1): 36-45.

[59] Bhandari AK (2011) On the distribution of estimated technical efficiency in stochastic frontier models: revisited, 10(1): 69-80.

[60] Bogetoft P, Nielsen K (2005) Internet based benchmarking. GDN, 14 (3): 195-215.

[61] Zhu J (2000) Multi-factor performance measure model with an application to Fortune 500 companies. European Journal of Operational Research, 23: 105-124.

[62] Zhu J (2004) Quantitative models for performance evaluation and benchmarking - data envelopment analysis with spreadsheets and DEA excel solver. Boston, USA: Kluwer Academic Publishers.

[63] Seiford LM, Zhu J (2003) Context-dependent data envelopment analysis - measuring attractiveness and progress. Omega, 31: 397-408.

[64] Ulucan A, Atici KB (2010) Efficiency evaluations with context-dependent and measure-specific data envelopment approaches: an application in a world bank supported project. Omega, 2010, 38:68-83.

[65] Johnson SC (1967) Hierarchical clustering schemes. Psychometrika, 32 (3): 241-254.

[66] MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In 'Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability', University of California Press, 281-297.

[67] McLachlan GJ, Peel D (2000) Finite mixture model. John Wiley & Sons.

[68] Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97 (458): 611-631.

[69] Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, Springer Science + Business Media, 520-534.

[70] Ward JH (1963) Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58 (301): 234-244.

[71] Mount DW (2004) Bioinformatics: sequence and genome analysis (2nd edition). John Inglis, 638.

[72] McGarigal K, Cushman S, Stafford SG (2000) Multivariate statistics for wildlife and ecology research. Springer-Verlag, 99.

[73] Causton HC, Quackenbush J, Brazma A (2003) Microarray/gene expressions data analysis: a beginner's guide. Blackwell Science.

[74] Russell D (1989) The principles of computer networking. Cambridge University Press, 38.

[75] Steinbach M, Karypis G, Kumar V (2000) A comparison of document clustering techniques. In KDD Workshop on Text Mining.

[76] Ovaska K, Laakso M, Hautaniemi S (2008) Fast gene ontology based clustering for microarray experiments. BioData Min, 1 (1): 11.

[77] Zhong S, Ghosh J (2003) A unified framework for model-based clustering. Journal of Machine Learning Research, 4: 1001-1037.

[78] Pelleg D, Moore A (2000) X-means: extending K-means with efficient estimation of the number of clusters. In 'Proceedings of the seventeenth international conference on machine learning', 727-734.

[79] Melin P, Castillo O (2005) Hybrid intelligent systems for pattern recognition using soft computing: an evolutionary approach for neural networks and fuzzy systems. Springer-Verlag, 172-178.

[80] Heyer LJ, Kruglyak S, Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. Genome Research, 9(11): 1106-1115.

[81] Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley.

[82] Medvedovic M, Sivaganesan S (2002) Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics, 18 (9): 1194-1206.

[83] Vaithyanathan S, Dom B (2003) Model-based hierarchical clustering. In Proceeding of the 16th Conference on Uncertainty in Artificial Intelligence (IPDPS'03), Morgan Kaufmann Publishers, 599-608.

[84] Kohonen T (1997) Self-organizing map. Springer-Verlag.

[85] Smyth P (2000) Model selection for probabilistic clustering using cross-validated likelihood. Statistics and Computing, 9 (1): 63-72.

[86] Ji Y, Wu C, Liu P, Wang J, Coombes RK (2005) Applications of beta-mixture models in bioinformatics. Bioinformatics, 21 (9): 2118-2122.

[87] Aigner H (1974) A new look at the statistical identification model. IEEE Transactions on Automatic Control, 19: 716-723.

[88] Schwarz G (1978) Estimating the dimension of a model. Annals of Statistics, 6 (2): 461-464.

[89] Bozdogan H (1987) Model selection and Akaike Information Criterion (AIC): the general theory and its analytic extensions. Psychometrika, 52 (3): 345-370.

[90] Biernacki C, Govaert G (1999) Choosing models in model-based clustering and discriminant analysis. Journal of Statistical Computation and Simulation, 64 (1): 49-71.

[91] Pan W (2006) Incorporating gene functions as priors in model-based clustering of microarray gene expression data, Bioinformatics, 22 (7): 795-801.

[92] Kumbhakar S C, Orea L (2004) Efficiency measurement using a latent class stochastic frontier model, Empirical Economics, doi: 10.1007/s00181-003-0184-2.

Publication I

Xiaofeng Dai. Non-parametric efficiency estimation using Richardson-Lucy blind deconvolution. *European Journal of Operational Research*, 248, 731-739, 2016.

© 2016 Elsevier Publications.
Reprinted with permission.

European Journal of Operational Research 248 (2016) 731–739

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Interfaces with other disciplines

Non-parametric efficiency estimation using Richardson–Lucy blind deconvolution

Xiaofeng Dai [a,b,*]

[a] Department of Information and Service Economy, Aalto University, 00101 Helsinki, Finland
[b] School of Biotechnology, JiangNan University, 214122 Wuxi, China

ARTICLE INFO

Article history:
Received 25 November 2013
Accepted 5 August 2015
Available online 15 August 2015

Keywords:
Stochastic frontier estimation
Richardson–Lucy blind deconvolution
Efficiency estimation
Nonparametric

ABSTRACT

We propose a non-parametric, three-stage strategy for efficiency estimation in which the Richardson–Lucy blind deconvolution algorithm is used to identify firm-specific inefficiencies from the residuals corrected for the expected inefficiency μ. The performance of the proposed algorithm is evaluated against the method of moments under 16 scenarios assuming $\mu = 0$. The results show that the Richardson–Lucy blind deconvolution method does not generate null or zero values due to wrong skewness or low kurtosis of inefficiency distribution, that it is insensitive to the distributional assumptions, and that it is robust to data noise levels and heteroscedasticity. We apply the Richardson–Lucy blind deconvolution method to Finnish electricity distribution network data sets, and we provide estimates for efficiencies that are otherwise inestimable when using the method of moments and correct ranks of firms with similar efficiency scores.

© 2015 Elsevier B.V. and Association of European Operational Research Societies (EURO) within the International Federation of Operational Research Societies (IFORS). All rights reserved.

1. Introduction

Productive efficiency analysis, which is a quantitative approach for evaluating the performance of a firm, can, e.g., offer insights into its performance and help managers make correct decisions. Generally, a productive efficiency analysis can be viewed as a two-step process: first, the production or cost frontiers are estimated by using parametric or non-parametric methods, and then the inefficiencies from the residuals estimated in the first step are predicted. In neoclassical theory based approaches such as Data Envelopment Analysis (DEA), the residuals are considered to be the inefficiencies (Charnes, Cooper, & Rhodes 1978; Farrell, 1957). In frontier production models such as Stochastic Frontier Analysis (SFA) (Aigner, Lovell, & Schmidt 1977; Meeusen and Van den Broeck (1977)) and the stochastic semi-parametric model Stochastic Non-smooth Envelopment of Data (StoNED) (Kuosmanen & Kortelainen, 2012), the residuals are assumed to be a composite of both the inefficiencies and the random noise. In the single output multiple input setting, StoNED contains the traditional DEA and SFA as its special cases. In stochastic frontier models, two-stage strategies are conventionally used for efficiency estimation, wherein the conditional mean $E(y|\mathbf{x})$ or the frontier is estimated in the first stage and the disturbance term (difference between the estimated and observed y) is decomposed into the inefficiency and the random noise in the second stage. In the first stage, the frontier can be estimated by using parametric or nonparametric regression techniques. Parametric models postulate a specific functional form for f and the parameters are estimated using techniques such as Modified Ordinary Least Squares (MOLS) (Aigner et al. 1977) and the maximum likelihood (ML) approach, with the latter being more frequently used. Non-parametric models do not assume a particular functional form but they do need to satisfy certain regularity axioms, with the frontier being determined using, e.g., Convex Nonparametric Least Squares (CNLS) (Kuosmanen & Kortelainen 2012). Keshvari and Kuosmanen (2013) relaxed the concavity assumption of CNLS (Keshvari & Kuosmanen, 2013). Given that semi-parametric approaches such as StoNED bridge the gap between DEA and SFA, there is a growing interest in this method. StoNED is a well-established method that is superior to other existing methods given its stochastic and semi-parametric nature (Kuosmanen & Kortelainen, 2012). By adopting the StoNED framework and relaxing its parametric assumptions, a fully nonparametric approach for efficiency identification that integrates the standard DEA and SFA models can be developed. In StoNED, techniques such as method of moments (MM) are conventionally used for identifying efficiencies from the residuals coming from the first step. Wang et al. (2014) developed a quantile-version of CNLS and StoNED (Wang, Wang, Dang, & Ge, 2004). However, these methods heavily depend on the accuracy of the distributional assumption of the error components and thus suffer from many problems such as wrong skewness (Kuosmanen & Fosgerau,

* School of Biotechnology, JiangNan University, 214122 Wuxi, China. Tel.: +8618611479958.
E-mail address: xiaofeng.dai@me.com

http://dx.doi.org/10.1016/j.ejor.2015.08.004
0377-2217/© 2015 Elsevier B.V. and Association of European Operational Research Societies (EURO) within the International Federation of Operational Research Societies (IFORS). All rights reserved.

2009). In addition to the two-stage strategies, other approaches have been applied in stochastic frontier models to account for the impact of environmental factors. The frontier inefficiency residuals were modeled as a function of various causal factors and a random component to study the systematic effect of the conditions that contribute to inefficiencies (Reifschneider & Stevenson, 1991). A generalized production frontier approach was reported by (Kumbhakar, Ghosh, & McGuckin, 1991) to estimate the determinants of inefficiencies. Huang and Liu proposed a hybrid of a stochastic frontier regression: the model combines a stochastic frontier regression and a truncated regression to estimate the production frontier with non-neutral shifting of the average production function (Huang & Liu, 1994). Conditional efficiency measures, such as conditional FDH, conditional DEA, conditional order-m and conditional order-α, have rapidly developed into a useful tool to explore the impact of exogenous factors on the performance of DMUs in a nonparametric framework (Daraio & Simar, 2005, 2007a, 2007b). A more recent paper examined the impact of environmental factors on the production process in a new two-stage type approach by using conditional measures to avoid the drawbacks of the traditional two-stage analysis, which provides a measure of inefficiency whitened from the main effect of the environmental factors (Badin, Daraio, & Simar, 2012).

Deconvolution has previously been shown to be a useful statistical technique for unknown density recovery (Meister 2006) which, in most cases, requires specifying the measurement error distribution (Stefanski & Carroll, 1990). For example, Kneip et al. (2012) applied deconvolution to estimate the boundary of a production set foe which the measurement error has an unknown variance; however, a lognormal distribution of the noise term is crucial to ensure the identifiability in context (Kneip, Simar, & Van Keilegom, 2012). Additionally, Schwarz et al. (2010) defined an estimator of the frontier function where partial information on the error distribution was assumed, i.e., zero-mean Gaussian random variable with an unknown variance (Schwarz, Van Bellegem, & Florens, 2010). Meister (2006) relaxed this assumption and consistently estimated both the target density and the unknown variance of the normal error, assuming that the target density was from the ordinary smooth family of distributions (Meister, 2006). Although fewer assumptions were needed for the error term in Meister's estimator, the target distribution was restricted to distributions such as Laplace, exponential, and gamma. Other attempts at relaxing constraints were made under a scenario wherein the contaminated errors ε ($\varepsilon = u + v$, u and v of each stand for the inefficiency and random noise, respectively, were not directly observable but represented an additive term of a regression model such as $y = \alpha + \beta\mathbf{x} + \varepsilon$ (α and β are the coefficients; $\mathbf{x}$ and y are the inputs and output). Horowitz and Markatou (1996) developed an estimator to handle cases that do not require specifying the component distributions of ε (Horowitz & Markatou, 1996). However, this method relies on the information along the time-axis of the panel data to identify densities in the composite error term, which cannot be applied to cross-sectional data, whose the error density is rarely entirely known. More importantly, Horrace and Parmeter (2011) proposed a cross-section complement of Horowitz and Markatou's method, which proved to be semi-uniformly consistent in identifying target density if u is ordinary smooth (Horrace & Parmeter, 2011). As a regression generalization of (Meister, 2006), the constraints posed in Meister's estimator are inherited in this method. For example, it is semi-parametric because it relies on a distributional law for v and because the density of u belongs to the ordinary smooth family. Further, as the methods of (Horowitz & Markatou, 1996) and (Horrace & Parmeter, 2011) work for data of the regression form, replacing ε with the regression residuals may introduce frontier estimation errors and can thus lead to a biased estimation of the inefficiencies.

Unlike the aforementioned efforts for applying deconvolution in frontier estimation, we are interested in inefficiency estimation using deconvolution in a non-parametric stochastic setting. To overcome the difficulty of estimating the expected inefficiency using kernel deconvolution, we return to the field where deconvolution is originated and explore the existing techniques. Deconvolution was originally applied in signal and image processing, where the point spread function (PSF) is used to describe the response of an imaging system to a point source (Haykin, 1993). Projected onto efficiency estimation problems, it is equivalent to the function of converting the inefficiencies to the observed residuals. Blind deconvolution is a technique for recovering the blurred object without any prior knowledge of the PSF (which is often costly or impossible to obtain). There are five categories of blind deconvolution methods: *a priori* blur identification methods (Cannon, 1976), zero sheet separation methods (Ghiglia, Romero, & Mastin, 1993), autoregressive moving average (ARMA) parameter estimation methods (Biemond, Tekalp, & Lagendijk, 1990), nonparametric methods based on high-order statistics (HOS) (Jacovitti & Neri, 1990; Wu, 1990), and nonparametric iterative methods (Ayers & Dainty, 1988; Kundur & Hatzinakos, 1998; McCallum, 1990). These methods differ in their assumptions about the PSF and the true object. After considering the advantages and limitations of each method, we are left with the nonparametric iterative methods. We restrict our options in this fashion because (1) the *a priori* methods are parametric; (2) zero sheet separation methods are highly sensitive to noise and prone to inaccuracy for large objects; (3) the ARMA parameter estimation methods may converge poorly and be computationally expensive if the number of parameters is very large; (4) nonparametric methods based on HOS require accurate modeling of the true object by a known non-Gaussian probability distribution and may be trapped in local minima in the estimation process; and (5) the results from algorithms in the first three categories are usually not unique unless additional assumptions are made about the PSF. Nonparametric iterative methods iteratively estimate PSF and the true object without any prior parametric assumptions. However, several constraints are required that, in the context of efficiency analysis, are as follows: (1) the inefficiencies are non-negative, (2) the range of inefficiency is known (e.g., within 0 and 1), and (3) the background noise is random. Typical algorithms that belong to this class are non-negativity and support constraints recursive inverse filtering (NAS–RIF) (Kundur & Hatzinakos, 1998), simulated annealing (SA) (McCallum, 1990), and iterative blind deconvolution (IBD) (Ayers & Dainty, 1988), which differ in their objectives of minimizing the cost functions and how these functions are constructed. Because NAS–RIF has certain requirements for the PSF, such as bounded-input bounded-output (BIBO), and because the choice of iteration parameters (e.g., perturbation scale) in SA is difficult, which affects its performance and convergence rate, we return our focus interest to IBD. IBD minimizes the cost function with respect to both the PSF and the true object simultaneously, and it is the most widely applied algorithm in blind deconvolution. The typical algorithms adopted for IBD in its iterative operations include a Wiener-type filter or the Richardson–Lucy (RL) algorithm. Because the Wiener-type filter assumes stationary noise, we are left with the RL algorithm. Further appealing is the probabilistic nature of the RL algorithm. We thus chose to apply the blind RL deconvolution (Fish, Brinicombe, & Pike, 1995) algorithm (abbreviated as RLb here) for inefficiency estimation.

The performance of the RLb was tested against that of MM (which is conventionally used in StoNED) under sixteen simulated scenarios, including those from (Aigner et al., 1977). In the RL deconvolution algorithm, the true object (e.g., inefficiency) was assumed to follow a Poisson distribution. By approximating a Poisson distribution using a Gaussian distribution which was assumed for the inefficiency term, we added a sufficiently large term to the inputs and subtracted it from the deconvoluted results. The results show that the RLb method outweighs MM in at least 4 aspects. It is (1) nonparametric and exempted from any distributional assumption, which

leads to (2) the circumvention of many common issues such as the wrong skewness problem; (3) it is insensitive to data noise; and (4) it is robust to data heteroscedasticity. Additionally, we applied the RLb method to an empirical problem, which used the residuals taken from the cost frontier estimation of 89 Finnish electricity distributors. We are among the pioneers in deploying deconvolution in efficiency estimation, and we are the first to identify inefficiencies using a fully non-parametric method.

The rest of the text is organized as below. The non-parametric three-stage efficiency estimation procedure wherein RLb is used for firm-specific inefficiency decomposition is described in detail in the 'Methods'. In the 'Monte Carlo Simulation', the data generating processes, performance measures and results are described and summarized. The data and results of the real case application are presented and discussed in the 'Empirical study' section. Finally, we summarize the key findings, contributions, limitations and possible future directions in 'Conclusion'.

2. Methods

2.1. Three-stage efficiency estimation

Considering the stochastic frontier model, let y_i represent the output of firm i, F denote the production function characterising the technology, $\mathbf{x} \in \Re_+^m$ being the vector of inputs and ε_i showing the composite errors, the standard econometric production model could be written as (1),

$$y_i = F(\mathbf{x}_i) + \varepsilon_i, \tag{1}$$

where the disturbance term ε_i of firm i could be decomposed into inefficiency u_i and random noise v_i, i.e., $\varepsilon_i = v_i - u_i$. The inefficiency term is composed of two parts, i.e., the expected inefficiency which is the same across firms μ and the firm-specific inefficiencies u_i. Thus, $u_i = \mu + u_i$, and $\varepsilon_i = v_i - \mu - u_i$ $(\mu + u_i \geq 0)$.

We propose a three-stage strategy to estimate inefficiencies using the blind Richardson–Lucy deconvolution (Fish et al., 1995) algorithm (abbreviated as RLb). The core of this strategy is Stage 3 where the RLb algorithm is used to decompose firm-specific inefficiencies from the corrected composite errors, which is independent of the form of the frontier and how the frontier is estimated. We employ CNLS regression for frontier estimation given its non-parametric nature. According to the duality theory, the production technology can be equivalently modeled by, e.g., the cost function (Kuosmanen, 2008), allowing the application of this algorithm to a wide range of problems. However, our model may not be feasible for problems with multiput-output and no cost minimization assumption, such as public sector organizations. Here, we use the production frontier model to illustrate this strategy.

- Stage 1: Estimate the shape of function F by CNLS regression and obtain the residuals ε_i, where the model is defined as (1) and F has no particular functional form but satisfies monotonicity and concavity.
- Stage 2: Estimate μ for all firms, and correct CNLS residuals by μ, i.e., $\epsilon_i = \varepsilon_i + \mu$, where ϵ_i is the corrected CNLS residual of firm i.
- Stage 3: Estimate the firm-specific inefficiencies using RLb, provided with the corrected CNLS residuals ϵ_i.

In this model, the inefficiency term is assumed to be comprised of the expected inefficiency shared among all firms (μ) and firm-specific inefficiencies u_i (i refers to firm i). The term μ was not captured using conventional methods which comprises of, e.g., technological bottleneck, economic environment, government regulation, etc. On the other hand, μ could be considered as part of the frontier, which is identified in Stage 2 and used for frontier correction. The term u_i contains the inefficiencies we are interested to identify which reflects the differences among firms.

2.1.1. Stage 1: CNLS regression

Both parametric and non-parametric models could be used for frontier estimation in the first stage. It can be analytically represented by (2) to (5),

$$\min_{\boldsymbol{\varepsilon},\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{i=1}^{N} \varepsilon_i^2 \quad \text{such that} \tag{2}$$

$$y_i = \alpha_i + \boldsymbol{\beta}_i^{'} \mathbf{x}_i + \varepsilon_i \tag{3}$$

$$\alpha_i + \boldsymbol{\beta}_i^{'} \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}_h^{'} \mathbf{x}_i \, \forall h, i = 1 \ldots N \tag{4}$$

$$\boldsymbol{\beta}_i \geq \mathbf{0} \, \forall i = 1 \ldots N \tag{5}$$

where α_i and $\boldsymbol{\beta}_i$ are coefficients specific to observation i, v_i $(v_i = \varepsilon_i + \mu - u_i)$ captures its random noise, and $\mathbf{x}_i$ is the vector of inputs for firm i.

For the CNLS estimator, the coefficients $\hat{\alpha}_i$, $\boldsymbol{\beta}_i$ obtained as the optimal solution to (2) to (5) are not necessarily unique. Denote the family of alternate optima as $F^{\star}$, the non-uniqueness issue is addressed by the following lower bound

$$\hat{g}_{\min}(\mathbf{x}) = \min_{\alpha \in \Re, \boldsymbol{\beta} \in \Re_+^m} \{\alpha + \boldsymbol{\beta}^{'} \mathbf{x} | \alpha + \boldsymbol{\beta}^{'} \mathbf{x}_i \geq \hat{y}_i, \forall i = 1, ..., N\} \tag{6}$$

Specifically, $\hat{g}_{\min}$ is the tightest lower bound for the family of functions $F^{\star}$. Note that for the observed data points $\mathbf{x}_i$, the fitted values are always unique, i.e., $g(\mathbf{x}_i) = \hat{g}_{\min}(\mathbf{x}_i), \forall i = 1, \ldots, N$.

2.1.2. Stage 2: Hall and Simar method

Many algorithms could be used in the second stage for μ estimation. For example, Hall and Simar have proposed a nonparametric method for estimating μ (abbreviated as HS) based on the unknown density of the composite error term (Hall & Simar, 2002), which could be coupled with RLb for efficiency estimation nonparametrically. In HS, the kernel density estimator is used for estimating the density function f, i.e.,

$$\hat{f}_{\varepsilon}(e) = (Nh)^{-1} \sum_{i=1}^{N} K\left(\frac{e - \varepsilon_i}{h}\right) \tag{7}$$

where $K(\cdot)$ is a compactly supported kernel, h is a bandwidth, ε is the composite error and e is the projection of ε on a line (i.e., the projected data of ε). Hall and Simar (2002) show that in the neighborhood of μ, the first derivative of the density function of the composite error term $(f_{\varepsilon}^{'})$ is proportional to that of the density function of the inefficiency term $(f_u^{'})$ (Hall & Simar, 2002). Due to the assumption that f_u has a jump discontinuity at 0, the CNLS residual ε has a jump discontinuity at $-\mu$ (Hall & Simar, 2002). Therefore, $\hat{\mu} = \arg\min_{\varepsilon \in \zeta} (\hat{f}_{\varepsilon}^{'}(\varepsilon))$ provides a nonparametric estimator of μ, where ζ is a closed interval in the right tail of $f_{\varepsilon}(\cdot)$. To implement HS, a bandwidth must be chosen and ζ need to be defined. According to (Delaigle & Gijbels, 2004), the following iterative procedure could be adopted to obtain $\hat{\mu}$.

- *Step 1* (Initialize h and ζ): Initialize the bandwidth by $h_0 \in CN^{-\frac{1}{9}}$, where C is a large number, e.g., 10, and $\zeta_0 = [\max \varepsilon_i^0 - h, \max \varepsilon_i^0]$.
- *Step 2* (Estimate μ): Estimate $\hat{\mu}_0$ using h_0 and ζ_0.
- *Step 3* (Update h and ζ): Set $h_1 = 0.85h_0$ and $\zeta_1 = [\hat{\mu}_0 - h_1, \hat{\mu}_0 + h_1]$, which are used to obtain $\hat{\mu}_1$.
- *Step 4* (Iteration and stop): Repeat steps 2 and 3 by $h_l = 0.85h_{l-1}$ and $\zeta_l = [\hat{\mu}_{l-1} - h_l, \hat{\mu}_{l-1} + h_l]$ where l is the index of this iteration. Stop the process when $|\hat{\mu}_l - \hat{\mu}_{l-1}| \leq N^{-\frac{2}{5}} |\hat{\mu}_1 - \hat{\mu}_0|$.

2.1.3. Stage 3: Richardson–Lucy blind deconvolution method

The RLb algorithm (the blind form of the RL algorithm), a nonparametric approach, is proposed here in the third stage to estimate firm-specific inefficiencies if μ is adjusted in the residuals, i.e., $\epsilon_i = \varepsilon_i + \mu$. The RL algorithm is originally developed for image recovery. According to (Richardson, 1972), given the blurred image B and the clear image I, the intensity I_p at the pixel location p is computed from the pixel intensities B_q by $P(I_p) = \sum_q P(I_p|B_q)P(B_q)$ where $P(I_p)$ can be identified as the distribution of I_p and so forth. Expanding $P(I_p|B_q)$ by Bayes's rule, $P(I_p) = \sum_q \frac{P(B_q|I_p)P(I_p)}{\sum_z P(B_q|I_z)P(I_z)} P(B_q)$. The best of a bad situation is used to break the dependency of $P(I_p)$ on both sides, where the current estimation of $P(I_p)$ is used to approximate $P(I_p|B_q)$. Thus,

$$P^{j+1}(I_p) = \sum_q \frac{P(B_q|I_p)P^j(I_p)}{\sum_z P(B_q|I_z)P^j(I_z)} P(B_q) = P^j(I_p) \sum_q P(B_q|I_p) \frac{P(B_q)}{\sum_z P(B_q|I_z)P^j(I_z)}, \quad (8)$$

where j is the index of the RL iteration.

Considering $B' = \sum_z P(B_q|I_z)P^j(I_z)$ to be the predicted blurry image according to the current estimation of clear image I^j (a more workable notation for $P^j(I_z)$), define $P(B_q|I_z) = PSF(q,z)$, and use $E_q = \frac{B_q}{B'_q}$ to denote the residual errors between the real and predicted blurry image, we get $\sum_q P(B_q|I_p)\frac{P(B_q)}{\sum_z P(B_q|I_z)P^j(I_z)} = \sum_q P(B_q|I_p)E_q^j$. If the isoplanatic condition holds, i.e., PSF is spatially invariant or $PSF(q,z)$ is the same for all q, $B' = \sum_z P(B_q|I_z)P^j(I_z) = I^j \otimes PSF$, and $\Sigma_q P(B_q|I_p)E_q$ becomes $PSF \star E_q$, where $\star$ and $\otimes$ are the correlation and convolution operators, respectively (note that the summation index in the generation of predicted blurry image, B', is z, and that for the integration of errors, E, is q). Hence, (8) becomes $I^{j+1} = I^j \times PSF \star \frac{B}{I^j \otimes PSF} = I^j \times PSF \star E^j$, where $E^j = \frac{B}{I^j \otimes PSF}$. In a two-dimensional space, this isoplanatic condition implies a symmetry condition in the positive region. Although such an assumption may introduce bias when the inefficiency distribution is asymmetric, it circumvents issues raised by the asymmetric assumption such as the wrong skewness problem and improves the estimation accuracy regarding the rankings.

In the context of inefficiency estimation, the inefficiency u and the residual ϵ could be identified as the clear image I and the blurry image B, respectively, and the noise v could be modelled as PSF. Thus, the iterative RL algorithm could be reformed as

$$u_i^{j+1} = u_i^j \times v \star \frac{\epsilon_i}{u_i^j \otimes v_i}, \quad (9)$$

The inefficiency estimation problem can be viewed as a projection of the image processing problem from the three dimension to a two-dimensional space. In image processing, the disturbance of pixel i on pixel j is dependent on the distance measuring their physical locations, while such disturbance is dependent on the distance measuring the similarities between the operational and managerial structures of the firms in inefficiency estimation. The iid (independent and identically distributed) condition is traditionally assumed using conventional efficiency estimation methods, while no particular assumption is needed when RL is used, assuring its accuracy and applicability.

In the blind form of the RL algorithm, PSF (i.e., v here) is unknown and is iteratively estimated together with u. Let m be the index of the blind iteration, j be the index of the Richardson–Lucy iteration, and i be the index of firms, the iterative estimation procedure of the RLb algorithm is summarized step by step below and illustrated in Fig. 1. The iterative process endows RLb the 'blindness' which assures its non-parametric property.

- Step 1: Initialize $v_0 = \mathbf{1}$ and $u_0 = \boldsymbol{\epsilon}$
- Step 2: For the mth blind iteration, do the following RL iteration steps until convergence:

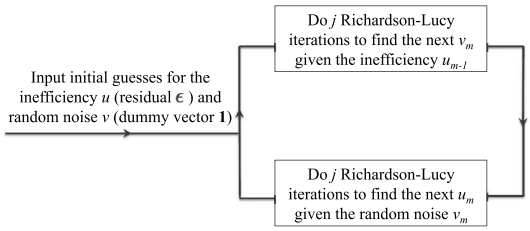


Fig. 1. Illustration of the Richardson-Lucy blind deconvolution algorithm.

  - Step 2.1: Estimate v for a specified number $(j+1)$ of RL iterations: do the jth RL iteration to find v for $j+1$ iterations, i.e., v_m^{j+1}, assuming u is known from the $m-1$ iteration.

$$v_m^{j+1} = v_m^j \times u_{m-1} \star \frac{\epsilon_i}{u_{i,m-1} \otimes v_m^j} \quad (10)$$

  - Step 2.2: Estimate u for the same number $(j+1)$ of RL iterations: do the jth RL iteration to find u for $j+1$ iterations, i.e., u_m^{j+1}, given that v_m is evaluated from the full iteration of (10).

$$u_{i,m}^{j+1} = u_{i,m}^j \times v_m \star \frac{\epsilon_i}{u_{i,m}^j \otimes v_m} \quad (11)$$

- Step 3: Iterate the blind iterations until convergence.

The RLb algorithm (or RL) minimizes the difference between the original and predicted degraded signals, i.e., $\arg\min_j (\epsilon_i - \hat{\epsilon}_i)$, per pixel with convergence proven in (Irani & Peleg, 1991; Lucy, 1974), allowing it to identify the optimal inefficiency at each single point (e.g., for each firm at each time point in a panel setting). However, this does not guarantee it to find the global minimum if the frontier function is not convex. In inefficiency estimation, the frontier is most often either parametrically determined or non-parametrically constrained to be convex, resulting in a convex function ϵ_i. Under this context, RLb is guaranteed to find the global optimal at each data point. In cases where the frontier is non-parametrically determined and non-convex, additional techniques are needed to form an improved version of RLb which will be the next step. An initial guess of v_0 is required to start the algorithm, which is specified as 1's with the same dimension as the residuals here.

The RLb algorithm requires non-negative inputs. For which, we add a large enough positive constant M to the inputs and subtract the deconvoluted M afterwards. Note that M is an arbitrary constant which does not have any particular meaning, so as to the deconvoluted M. The results are independent of the choice of M as tested by simulations. Mathematically, the property of translation invariance is shown below:

$$\because \epsilon(\mathbf{x}_i + M) = u(\mathbf{x}_i + M) \otimes v, M > 0 \quad (12)$$

$$= u(\mathbf{x}_i) \otimes v + u(M) \otimes v \quad (13)$$

$$= \epsilon(\mathbf{x}_i) + \epsilon(M) \quad (14)$$

$$\therefore \epsilon(\mathbf{x}_i) = \epsilon(\mathbf{x}_i + M) - \epsilon(M) \quad (15)$$

$$u(\mathbf{x}_i) = u(\mathbf{x}_i + M) - u(M) \quad (16)$$

where $\mathbf{x}_i$ is the input of firm i, v is the noise, $\epsilon(\mathbf{x}_i)$ and $u(\mathbf{x}_i)$ are the corrected composite error and firm-specific inefficiency of firm i.

2.2. Performance comparison

The performance of RLb is compared with the efficiency decomposition method used in StoNED, i.e., method of moments (MM). The second and third stages are estimated together in MM. In particular, the variance parameters σ_u^2, σ_v^2 are estimated based on the skewness of the CNLS residuals obtained from Stage 1 with additional distributional assumptions, and the conditional expected values of inefficiencies are computed given the parameter estimates of σ_u^2 and σ_v^2 (details provided in Kuosmanen & Kortelainen, 2012). Thereby, the inefficiencies estimated from ε using RLb and MM are $u_{i,RLb} = u_{i,RLb} + \mu$ and $u_{i,MM}$, respectively. To make RLb and MM comparable and exclude the influence of other algorithms such as HS on performance evaluation, we assume $\mu = 0$ in the simulations and remove μ in the empirical study by (17), where $\mu = \bar{u}_{RLb} - \bar{u}_{RLb}$ and $\bar{u}_{RLb} = \bar{u}_{MM}$ (note that $\bar{v} = E(v) = 0$ leads to $\bar{u} = \bar{v} - \bar{\varepsilon} - \mu = -\bar{\varepsilon} - \mu$; thus $\bar{u}$ is independent of the estimation method, i.e., $\bar{u}_{RLb} = \bar{u}_{MM}$).

$$\begin{aligned} u_{i,RLb} &= u_{i,RLb} - \mu \\ &= u_{i,RLb} - (\bar{u}_{RLb} - \bar{u}_{RLb}) \\ &= u_{i,RLb} - (\bar{u}_{RLb} - \bar{u}_{MM}) \end{aligned} \tag{17}$$

3. Monte Carlo simulation

3.1. Data generating process

We designed two sets of simulations to assess the performance of the RLb method with a sample size of 100 for each simulation. The first set of simulations are analogous to (Aigner et al., 1977), with the scenarios designed for different signal to noise ratios ($\lambda = \frac{\sigma_u}{\sigma_v}$). The second simulation set is an extension of the first one, with the aim of testing the influence of different distributional assumptions on the inefficiency term and data heteroscedasticity on the estimation accuracy. Four distributions in addition to the half normal distribution, including three continuous densities conventionally assumed for the inefficiency term ('truncated normal', 'gamma', 'exponential' Kuosmanen & Kortelainen, 2012) and one discrete distribution ('Poisson'). The noise term was assumed to follow normal distribution, with zero mean and a variance of 0.3. The signal to noise ratio in the second simulation set was set to the middle value ($\lambda = 1.24$) of the first simulation set under all scenarios. Groupwise heteroscedasticity was generated for each heterscedastic data. Particularly, four equally divided sub-populations were generated, with consecutive data points being grouped together in their generic order (i.e., the first 25 data points belong to subgroup 1, points 26 to 50 belong to subgroup 2, and so on). The RLb method and MM (assuming half normal distribution for the inefficiency term) were applied to each scenario, with 100 iterations (Table 1).

Each scenario is given a four-digit name. The first letter is the initial of the inefficiency distribution, i.e., 'H', 'T', 'G', 'E', 'P' are short for the half normal, truncated normal, gamma, exponential and Poisson distribution, respectively. The second digit shows the signal to noise ratio, which is represented by λ and defined as $\lambda = \frac{\sigma_u}{\sigma_v}$, with 2, 1, 0 representing the high, moderate and low levels, respectively. Here, this statistic is taken from (Aigner et al., 1977), i.e., 2 is equivalent to $\lambda = 1.66$, 1 means $\lambda = 1.24$, and 0 is short for $\lambda = 0.83$. The third character indicates whether the data is heteroscedastic, where '−' and '+' each means without and with heteroscedasticity. The last number shows the constant $\bar{u}$ which is 0 here.

3.2. Performance measures

Mean squared error (MSE) was used to measure the performance of the algorithm, which is defined as

$$\mathrm{MSE}_{\mu_u} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}_{u_i} - \mu_{u_i})^2, \tag{18}$$

Table 1
Parameter settings in the simulations. In gamma distribution, $\mu_u = \theta k, \sigma_u = \theta\sqrt{k}$. In exponential distribution, $\mu_u = \sigma_u = \theta\sqrt{k}$. In Poisson distribution, $\sigma_u = \sqrt{\mu_u}$. Groupwise heteroskedasticity (four subgroups) is generated for 'H1+0', 'T1+0', 'G1+0', 'E1+0', 'P1+0' (consecutive data points are grouped together in their generic order, i.e., the first 25 data points belong to subgroup 1 and so on); the elements in the square brackets are the corresponding parameters in each subgroup; and if more than one parameter is needed for a particular distribution, the elements are ordered in the same way in the brackets for different parameters.

Scenario	Parameter setting	λ	Distribution
H2+0	$\mu_u = 0, \sigma_u = 0.8261$	1.66	Half normal
H1+0	$\mu_u = 0, \sigma_u = 0.6171$	1.24	Half normal
H0+0	$\mu_u = 0, \sigma_u = 0.4131$	0.83	Half normal
H1+0	$\mu_u = 0, \sigma_u = 0.6171$	1.24	Half normal
H1+0	$\mu_u = [0, 0, 0, 0]$, $\sigma_u = [0.4131, 0.5508, 0.6884, 0.8261]$	1.24	Half normal
T1+0	$\mu_u = 1, \sigma_u = 0.3882$	1.24	Truncated normal
T1+0	$\mu_u = [0.8, 1.2, 0.5, 1.5]$, $\sigma_u = [0.2598, 0.3465, 0.4331, 0.5197]$	1.24	Truncated normal
G1+0	$\theta_u = 1, k_u = 0.1384$	1.24	Gamma
G1+0	$\theta_u = [1, 1, 1, 1]$, $k_u = [0.0620, 0.1102, 0.1722, 0.2480]$	1.24	Gamma
E1+0	$\mu_u = 0.3720$	1.24	Exponential
E1+0	$\mu_u = [0.2490, 0.3320, 0.4150, 0.4980]$	1.24	Exponential
P1+0	$\mu_u = 0.1384$	1.24	Poisson
P1+0	$\mu_u = [0.0620, 0.1102, 0.1722, 0.2480]$	1.24	Poisson
All	$\mu_v = 0, \sigma_v = 0.3$, data size = 100, iterations = 100		

Table 2
Results of simulation set 1. Three scenarios are simulated in this set. In the scenario names, 'H2+0': $y = \varepsilon$, $\lambda = 1.66$; 'H1+0': $y = \varepsilon$, $\lambda = 1.24$; 'H0+0': $y = \varepsilon$, $\lambda = 0.83$. $\mu_v = 0$ and $\sigma_v = 0.3$ are used for data generation for all simulations, and 100 simulations are run for each scenario. 'TRUE', 'MSE', 'NUM' are the true value, minimum standard error and the number of the corresponding statistics. For MM method, only iterations with no NA and no zero values are used for statistics computing. 'Stat' and 'Met' are short for statistics and method, respectively. All statistics are rounded to 4 digits.

Type	Stat	Met	H2+0	H1+0	H0+0
TRUE	μ_u		0.6636	0.4945	0.3276
MSE	μ_u	RLb	0.0179	0.0176	0.0098
MSE	μ_u	MM	0.1462	0.0697	0.0177
TRUE	σ_u		0.4928	0.3675	0.2485
MSE	σ_u	RLb	0.005	0.0084	0.0114
MSE	σ_u	MM	0.0023	0.0035	0.0041
TRUE	λ		1.6774	1.2516	0.8459
MSE	λ	RLb	0.0567	0.0952	0.1293
MSE	λ	MM	0.0265	0.0389	0.0463
NUM	NA	MM	7	1	1
NUM	0	MM	1	2	18

$$\mathrm{MSE}_{\sigma_u} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}_{u_i} - \sigma_{u_i})^2, \tag{19}$$

$$\mathrm{MSE}_{\lambda} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_i - \lambda_i)^2, \tag{20}$$

where μ_u (mean of u), σ_u (standard deviation of u) and $\lambda = \frac{\sigma_u}{\sigma_v}$ (signal to noise ratio) are estimated over 100 iterations for each simulated scenario, and i is the index of the data points which is 100 in this study, i.e., N = 100. In addition, we counted the number of non-valid and zero estimates from both methods to assess their performance.

3.3. Results and discussions

The statistics of the two sets of simulation results are summarized in Tables 2 and 3. The estimated and true inefficiencies are plotted against each other in Figs. 2 and 3, where NA or zeros estimated using MM are excluded when drawing plots or computing the statistics.

Table 3
Results of simulation set 2. Ten scenarios are simulated in this set. In the scenario names, 'H1+0': half normal, homoscedastic; 'H1+0': half normal, heteroscedastic; 'T1+0': truncated normal, homoscedastic; 'T1+0': truncated normal, heteroscedastic; 'G1+0': gamma, homoscedastic; 'G1+0': gamma, heteroscedastic; 'E1+0': exponential, homoscedastic; 'E1+0': exponential, heteroscedastic; 'P1+0': Poisson, homoscedastic; 'P1+0': Poisson, heteroscedastic. $y = \varepsilon$, $\lambda = 1.24$, $\mu_v = 0$, $\sigma_v = 0.3$ are used for data generation for all simulations, and 100 simulations are run for each scenario. 'TRUE', 'MSE', 'NUM' are the true value, minimum standard error and the number of the corresponding statistics. For MM method, only iterations with no NA and no zero values are used for statistics computing. 'Stat' and 'Met' are short for statistics and method, respectively. All statistics are rounded to 4 digits.

Type	Stat	Met	H1+0	H1+0	T1+0	T1+0	G1+0	G1+0	E1+0	E1+0	P1+0	P1+0
TRUE	μ_u		0.4944	0.4934	1.0010	1.0172	0.1380	0.1536	0.3666	0.3694	0.1331	0.1436
MSE	μ_u	RLb	0.0145	0.0235	0.0008	0.0045	0.07783	0.0799	0.0325	0.0362	0.0683	0.0738
MSE	μ_u	MM	0.0481	0.1227	0.0981	0.0177	0.0347	0.0086	0.1293	0.1723	0.2680	0.1873
TRUE	σ_u		0.3699	0.3728	0.3852	0.3715	0.3508	0.3662	0.3626	0.3549	0.3533	0.3624
MSE	σ_u	RLb	0.0078	0.0087	0.0111	0.0121	0.0094	0.0114	0.0087	0.0087	0.0067	0.0071
MSE	σ_u	MM	0.0043	0.0073	0.0021	0.0164	0.0528	0.0637	0.0064	0.0098	0.0135	0.0152
TRUE	λ		1.2380	1.2431	1.2878	1.2407	1.1706	1.2201	1.2122	1.1847	1.1825	1.2095
MSE	λ	RLb	0.0853	0.0955	0.1225	0.1324	0.1041	0.1269	0.0950	0.0956	0.0727	0.0797
MSE	λ	MM	0.0468	0.0803	0.0241	0.1842	0.5932	0.7212	0.0706	0.1073	0.1513	0.1731
NUM	NA	MM	2	19	0	0	88	83	29	34	83	89
NUM	0	MM	3	0	49	6	0	1	1	0	0	0

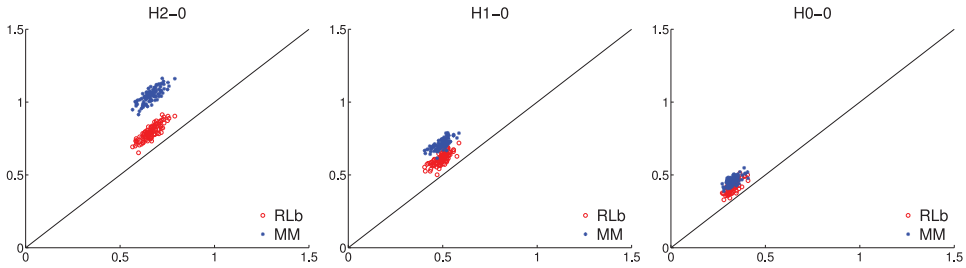


Fig. 2. Simulation set 1. 'H2+0': $\lambda = 1.66$, $y = \varepsilon$; 'H1+0': $\lambda = 1.24$, $y = \varepsilon$; 'H0+0': $\lambda = 0.83$, $y = \varepsilon$. Half normal distribution. For MM method, only iterations with no NA and no zero values are used for plotting.

3.3.1. RLb is robust to distribution skewness and kurtosis

As observed from Tables 2 and 3, no NA or zero value was produced using the RLb method, whereas in the case of MM such non-valid estimates were unavoidable and present throughout the simulations. The number of NAs increases with the level of inefficiencies (as represented by λ as σ_v is invariant here) using MM, which indicates the severity of the wrong skewness problem when estimating large inefficiencies using conventional methods such as MM. Such problems become worse when the distribution of the inefficiency is incorrectly assumed, e.g., almost 90% of the MM estimates are non-valid for gamma and Poisson distributional assumptions even with modest inefficiency levels (Table 3). In contrast, the number of zeros increases as the inefficiency decreases, which suggests of the poor performance of MM in identifying inefficiencies with low kurtosis. This problem is particularly severe under truncated normal assumption, with around half zero estimates being generated with modest λ (Table 3). These problems could be well circumvented by the RLb method, because of its 'blindness' to the inefficiency distribution as described below.

3.3.2. RLb is robust to distribution assumption

The RLb is insensitive to distribution assumptions, and it always produces better estimates than MM, even with the NA and zero values removed. The RLb method works particularly well when the distribution of inefficiencies is assumed to be truncated normal (Fig 3 (b and c)). Note that the RLb algorithm assumes isoplanatic conditions for both u and v which is analogous to the symmetric distributional assumption of u in a two-dimensional space. This leads to the less biased results in the case of a truncated normal distribution, where the density contains symmetric parts in the non-negative region. The MSE of the RLb estimates (μ_u) are larger than those from MM, which, however, has a much lower MSE of the standard deviation of inefficiencies (σ_u) and λ than does MM (Table 3). Given the isoplanatic constraints on u using RLb, this bias may vanish when the shape parameter $k > 1$ ($k = 1$ here). In the rest of our study cases, both RLb and MM overestimate the inefficiencies, though the estimates from RLb deviate less from the true values than do those from MM.

3.3.3. RLb is robust to data noise

The RLb method is more robust to data noise than MM. The distances between the RLb estimates and the true inefficiencies remain almost invariant, whereas those for MM changes linearly with λ. With σ_v staying invariant, λ increases with the level of inefficiencies. The divergence of the MM estimates from the true inefficiencies increases with the inefficiency level, which is greater than that of the RLb method even at the scenario with the lowest λ (Fig. 2(c)). When an explicit intercept term is extracted from the model ($y = 1 + \varepsilon$), the estimates from the RLb method are shifted by the intercept term, i.e., upwards by one here (Table 2), with no obvious change in σ_u or λ. These results indicate that the RLb method does not remove constant errors such as the difference between the estimated and true frontier ($f(x) - \hat{f}(x)$) and that the accuracy of inefficiency recovery depends on the frontier estimation accuracy. Additionally, these findings demonstrate the robustness of the RLb method to sources of data noise. Enlarging or shrinking the inefficiency term (assume $u^\star = u + 1$) by a constant does not considerably change the standard deviation of the estimates or the signal to noise ratio (λ). Note that when the distribution of inefficiencies is discrete (i.e., Poisson here), MM performs poorly, with outrageously large estimated standard deviations and a considerable overestimation of the inefficiencies observed here. Thereby, RLb outperforms MM in its consistent

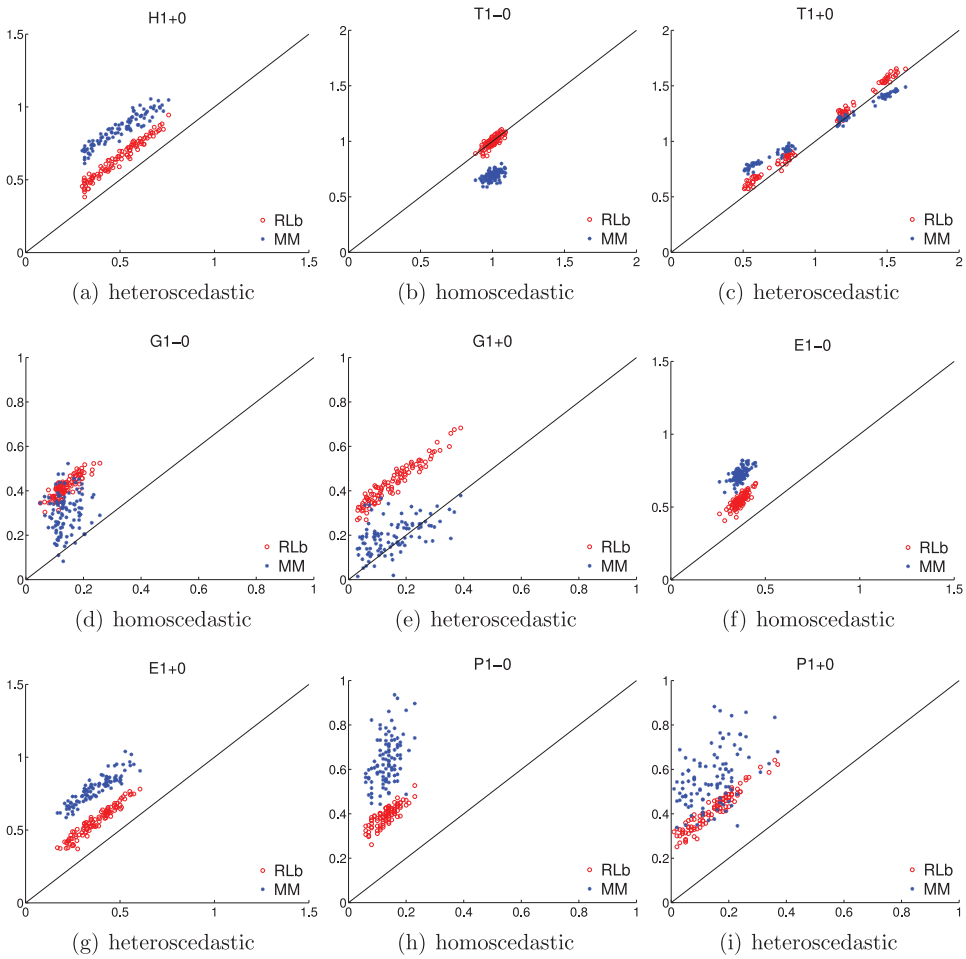


Fig. 3. Simulation set 2. 'H1+0': homoscedastic, half normal distribution, $y = \varepsilon$, $\lambda = 1.24$ (Fig 2b); 'H1+0': heteroscedastic, half normal distribution, $y = \varepsilon$, $\lambda = 1.24$ (a); 'T1+0': homoscedastic, truncated normal distribution, $y = \varepsilon$, $\lambda = 1.24$ (b); 'T1+0': heteroscedastic, truncated normal distribution, $y = \varepsilon$, $\lambda = 1.24$ (c); 'G1+0': homoscedastic, gamma distribution, $y = \varepsilon$, $\lambda = 1.24$ (d); 'G1+0': heteroscedastic, gamma distribution, $y = \varepsilon$, $\lambda = 1.24$ (e); 'E1+0': homoscedastic, exponential distribution, $y = \varepsilon$, $\lambda = 1.24$ (f); 'E1+0': heteroscedastic, exponential distribution, $y = \varepsilon$, $\lambda = 1.24$ (g); 'P1+0': homoscedastic, Poisson distribution, $y = \varepsilon$, $\lambda = 1.24$ (h); 'P1+0': heteroscedastic, Poisson distribution, $y = \varepsilon$, $\lambda = 1.24$ (i). For MM method, only iterations with no NA and no zero values are used for plotting.

estimation under various distributional assumptions of the inefficiency term. RLb can produce unbiased results when the inefficiency distribution has a symmetric part in the positive region, and it may overestimate the inefficiencies when such symmetry does not exist.

3.3.4. RLb is robust to data heteroscedasticity

The RLb method is insensitive to data heteroscedasticity. As observed from Table 3, the MSEs of μ_u, σ_u and λ are almost invariant for RLb estimates between data with and without heteroscedasticity, and these statistics are considerably larger for heterscedastic data than homoscedastic data for MM estimates.

4. Empirical study

To assess whether and to what extent the use of RLb instead of MM affects the results in a real world application, RLb is applied to Finnish electricity distribution networks, where the residuals from the cost frontier model is used for inefficiency estimation. The cost frontier model is defined in (21),

$$\ln x_i = \ln C(y_i) + \varepsilon_i, \tag{21}$$

where C represents the cost function and is modelled as a function of the output y, and ln means log e.

4.1. Data

The data consists of the six-year average over the period 2005–2010, which is available in the Energy Market Authority (EMA) website (www.emvi.fi). In the cost frontier model, the total cost (x) is used as the single input, three variables, i.e., 'Energy transmission' (GWh of 0.4 kV equivalents, y_1), 'Network length' (km, y_2), and 'Customer number' (y_3), are specified as the outputs ($\mathbf{y}$), and the proportion of the underground cables in the total network length is used as a contextual variable (z) to control the DMUs' heterogeneity and their operating environment. Specifically, x includes

Table 4
Descriptive statistics of the input, output, and contextual variables of the empirical data. 'MEAN', 'STD', 'MIN', 'MAX', 'KURT', 'SKEW' represent the 'Mean', 'Standard deviation', 'Minimum value', 'Maximum value', 'Kurtosis' and 'Skewness', respectively. The data are averaged over a six-year period 2005–2010.

Variable	MEAN	STD	MIN	MAX	KURT	SKEW
x = Total cost (1000€)	5052	10144	139	64326	22	4
y_1 = Energy transmission (GWh)	512	1026.65	15	6978	22	4
y_2 = Network length (km)	4370	10465.63	46	68349	26	5
y_3 = Customer number	37650	73856.08	24	426769	16	4
z = Underground cable proportion	0.23	0.28	0	1	0.43	1.27

Table 5
Results of empirical study. The 89 units are ordered by the rank of the efficiencies ('CE') estimated using RLb and listed in a column-wise fashion from left to right, i.e., the first 30 units are listed from top to bottom in the left column, and the last 29 units are listed from top to bottom in the right column. All statistics are rounded to four digits.

DMU	CE_{RLb}	CE_{MM}	DMU	CE_{RLb}	CE_{MM}	DMU	CE_{RLb}	CE_{MM}
32	1.2337	Inf	80	0.9793	0.9690	69	0.8787	0.9043
70	1.2108	Inf	62	0.9746	0.9660	3	0.8784	0.9041
22	1.2085	Inf	15	0.9723	0.9645	64	0.8754	0.9021
56	1.1429	Inf	31	0.9671	0.9612	87	0.8651	0.8953
37	1.1267	0.9990	63	0.9622	0.9582	89	0.8640	0.8947
59	1.1204	0.9990	58	0.9589	0.9561	1	0.8555	0.8890
28	1.1186	0.9987	79	0.9496	0.9501	55	0.8547	0.8885
38	1.0971	0.9986	33	0.9488	0.9496	48	0.8288	0.8712
57	1.0793	0.9983	60	0.9472	0.9486	43	0.8187	0.8645
50	1.0764	0.9983	71	0.9436	0.9463	67	0.8059	0.8558
61	1.0690	0.9980	76	0.9430	0.9460	53	0.8035	0.8542
73	1.0600	0.9976	35	0.9400	0.9440	78	0.8018	0.8530
39	1.0567	0.9974	81	0.9399	0.9440	25	0.7884	0.8439
46	1.0565	0.9974	30	0.9390	0.9434	17	0.7867	0.8428
49	1.0522	0.9971	7	0.9339	0.9401	36	0.7858	0.8422
83	1.0416	0.9961	40	0.9300	0.9376	8	0.7849	0.8415
75	1.0400	0.9959	4	0.9267	0.9355	29	0.7616	0.8255
45	1.0336	0.9950	77	0.9195	0.9309	5	0.7605	0.8247
6	1.0296	0.9942	41	0.9182	0.9300	54	0.7467	0.8152
82	1.0284	0.9939	42	0.9115	0.9257	13	0.7384	0.8094
24	1.0212	0.9919	19	0.9050	0.9214	14	0.7326	0.8053
16	1.0158	0.9900	85	0.9009	0.9188	9	0.7313	0.8045
44	1.0141	0.9892	26	0.8992	0.9177	34	0.7305	0.8039
20	1.0108	0.9877	2	0.8947	0.9147	27	0.7294	0.8031
10	1.0029	0.9835	86	0.8921	0.9131	51	0.7186	0.7955
74	1.0004	0.9821	72	0.8899	0.9116	18	0.7105	0.7898
66	0.9939	0.9781	11	0.8872	0.9098	23	0.7049	0.7858
12	0.9915	0.9766	52	0.8871	0.9098	65	0.6892	0.7746
68	0.9876	0.9742	84	0.8858	0.9090	88	0.6749	0.7643
21	0.9839	0.9719	47	0.8812	0.9059			

the operational expenditure and half of the interruption cost, and the electricity transmission at different voltage levels is weighted according to the average transmission cost such that lower weight is assigned to high-voltage transmission than low-voltage transmission in y_1. The descriptive statistics of the data are listed in Table 4, with more detailed description of the variables and the regulatory application available in (Kuosmanen, 2012).

4.2. Results and discussions

The cost efficiency score is estimated using $CE = \exp(u_i)$ for each unit i and summarized in Table 5. Except for the top 8 DMUs (according to RLb estimation), all firms are ranked in the same order using RLb and MM. The top 4 firms, i.e., 32, 70, 22 and 56, as ranked by RLb, have no CE score using MM estimation, because the inefficiencies are estimated to be infinite due to numerical problems. Using MM, the CE scores are rather close among the 4 DMUs, i.e., they range from 0.9990 to 0.9986, where the differences are more obvious using RLb, i.e., the DMUs range from 1.1267 to 1.0971. In this model, firm-specific inefficiencies are separated from the expectation of the inefficiency which is caused by environmental influences. A firm can balance such an adverse global impact by adopting superb technological or managerial renovations, thus exhibiting over-efficiencies in terms of firm-specific efficiency. The CE scores are plotted in Fig. 4, where CEs from RLb have larger amplitudes than MM estimates (excluding the infinite estimates), i.e., the standard deviation of the RLb estimates (0.1266) is nearly twice that of MM (0.0678), thus rendering the differences more distinguishable among DMUs. The average CE score is similar between RLb and MM (0.9263 for RLb, 0.9201 for MM), which indicates that the RLb estimates are firm-specific, with μ being adjusted and comparable to those from MM. Moreover, RLb is able to estimate efficiencies without numerical problems and to correctly distinguish firms that have similar CE scores. The sample correlation coefficient is close to 1 for CNLS and RLb inefficiency estimates, and it is 0.9903 when the inefficiencies are estimated using MM. Thus, the perfect correlation between the CNLS residuals and inefficiencies obtained using MM also applies to the RLb estimator. The key merits of RLb are that (1) it is non-parametric; (2) it faces no numerical problems with the outputs, such as values of NA; and (3) it can differentiate similar inefficiencies. Possessing these merits does not affect the principle properties of the outputs, e.g., the ranking order; therefore, a good correlation between inefficiencies obtained using RLb and MM (if not NA) is expected, which supports the precision of RLb given the wide applicability of MM.

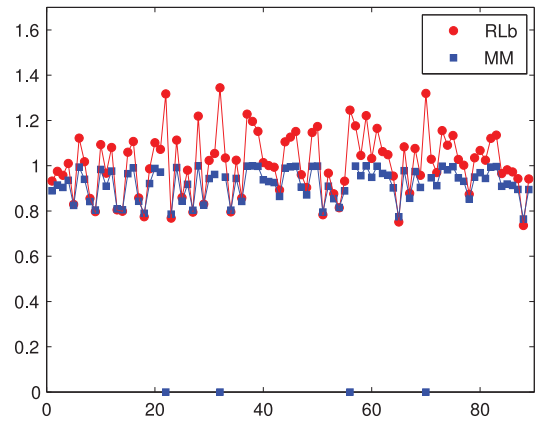


Fig. 4. Efficiencies estimated in the empirical study using RLb and MM. The x axis shows the number of the 89 electricity suppliers, and the y axis shows the efficiencies (CE). The firms without efficiencies shown as blue squared dots have infinite estimates.

5. Conclusions

In this study, we deploy a fully non-parametric algorithm, the RL blind deconvolution method, to decompose firm-specific inefficiencies from their composite errors corrected by the expected inefficiency $\mu = 0$ in productive efficiency analysis. By comparing the performance of RLb and MM under 13 scenarios assuming $\mu = 0$, we show that the RLb method outweighs conventional methods such as MM in four tested aspects. First, it never outputs null or zero values due to incorrect skewness or low kurtosis of the inefficiency density. Second, it is insensitive to the distributional assumption of the inefficiency term u, and it does not require any additional assumptions such as iid (independently and identically distributed) samples. Third, it is robust to data noise levels. Fourth, it gives consistent estimates, regardless of data heteroscedasticity. In addition, we applied RLb to the Finland electricity distribution network data set, wherein the efficiencies inestimable using MM are provided and firms with similar efficiency scores are correctly ranked. We are one of the pioneers in applying deconvolution in inefficiency estimation, and we are the first to report a fully non-parametric method for composite error decomposition, compared with other groups, which use kernel deconvolution techniques.

It is worth noting that the RLb algorithm was initially developed to solve image degradation problems in a three-dimensional space. Thus, its utility in panel data warrants further exploration. Additionally, we could extend RLb to solve cases wherein the frontier is non-convex and non-parametrically determined. Despite the advantages of RLb, we should be aware of its sensitivity to frontier estimation error, i.e., the inefficiency estimates are shifted by the difference between the estimated and true frontier. Additionally, the RLb method is not unbiased because of its isoplanatic assumption on u and v. Exploring how to overcome these problems and further improve the estimation accuracy are interesting topics next steps. In addition, applying this more robust tool to solve some empirical problems may offer high practical values and is suggested here for future research.

References

Aigner, D., Lovell, C. A., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics, 6*(1), 21–37.

Ayers, G. R., & Dainty, J. C. (1988). Iterative blind deconvolution method and its applications. *Optics Letters, 13*(7), 547–549.

Badin, L., Daraio, C., & Simar, L. (2012). How to measure the impact of environmental factors in a nonparametric production model. *European Journal of Operational Research, 223*, 818–833.

Biemond, J., Tekalp, A. M., & Lagendijk, R. L. (1990). Maximum likelihood image and blur identification: A unifying approach. *Optical Engineering, 29*(5), 422–435.

Cannon, M. (1976). Blind deconvolution of spatially invariant image blurs with phase. *IEEE Transactions on Acoustics Speech and Signal Processing, 24*(1), 58–63.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429–444.

Daraio, C., & Simar, L. (2005). Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of Productivity Analysis, 24*, 93–121.

Daraio, C., & Simar, L. (2007a). *Advanced robust and nonparametric methods in efficiency analysis. methodology and applications*. New York, USA: Springer.

Daraio, C., & Simar, L. (2007b). Conditional nonparametric frontier models for convex and non convex technologies: a unifying approach. *Journal of Productivity Analysis, 28*, 13–32.

Delaigle, A., & Gijbels, I. (2004). Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics & Data Analysis, 45*(2), 249–267.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society, 120*(3), 253–290.

Fish, D. A., Brinicombe, A. M., & Pike, E. R. (1995). Blind deconvolution by means of the richardson-lucalgorithm. *Journal of the Optical Society of America A, 12*(1), 58–65.

Ghiglia, D. C., Romero, L. A., & Mastin, G. A. (1993). Systematic approach to two-dimensional blind deconvolution by zero-sheet separation. *Journal of the Optical Society of America A, 10*(5), 1024–1036.

Hall, P., & Simar, L. (2002). Estimating a changepoint, boundary, or frontier in the presence of observation error. *Journal of the American Statistical Association, 97*(458), 523–534.

Haykin, S. (1993). *Blind deconvolution*. New York, USA: Prentice Hall.

Horowitz, J. L., & Markatou, M. (1996). Semiparametric estimation of regression models for panel data. *The Review of Economic Studies, 63*, 145–168.

Horrace, W. C., & Parmeter, C. F. (2011). Semiparametric deconvolution with unknown error variance. *Journal of Productivity Analysis, 35*, 129–141.

Huang, C. J., & Liu, J.-T. (1994). Estimation of a non-neutral stochastic frontier production function. *Journal of Productivity Analysis, 5*, 171–180.

Irani, M., & Peleg, S. (1991). Improving resolution by image registration. *Graphical Models and Image Processing, 53*(3), 231–239.

Jacovitti, G., & Neri, A. (1990). A bayesian approach to 2D non minimum phase AR identification. *Fifth ASSP Workshop on Spectrum Estimation and Modeling*, 79–83.

Keshvari, A., & Kuosmanen, T. (2013). Stochastic non-convex envelopment of data: applying isotonic regression to frontier estimation. *European Journal of Operational Research*, 49–481.

Kneip, A., Simar, L., & Van Keilegom, I. (2012). Boundary estimation in the presence of measurement error with unknown variance. *Discussion paper, ISBA, UCL, 2*. Replaces DP 1046

Kumbhakar, S. C., Ghosh, S., & McGuckin, J. T. (1991). A generalized production frontier approach for estimating determinants of inefficiency in us dairy farms. *Journal of Business and Economic Statistics, 9*, 286–297.

Kundur, D., & Hatzinakos, D. (1998). A novel blind deconvolution scheme for image restoration using recursive filtering. *IEEE Transactions on Signal Processing, 46*(2), 375–390.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *Econometrics Journal, 11*, 308–325.

Kuosmanen, T. (2012). Stochastic semi-nonparametric frontier estimation of electricity distribution networks: application of the StoNED method in the Finnish regulatory model. *Energy Economics, 34*, 2189–2199.

Kuosmanen, T., & Fosgerau, M. (2009). Neoclassical versus frontier production models? testing for the skewness of regression residuals. *The Scandinavian Journal of Economics, 111*(2), 351–367.

Kuosmanen, T., & Kortelainen, M. (2012). Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis, 38*(1), 11–28.

Lucy, L. B. (1974). An iterative technique for the rectification of observed distributions. *The Astronomical Journal, 79*(6), 745–754.

McCallum, B. C. (1990). Blind deconvolution by simulated annealing. *Optics Communications, 75*(2), 101–105.

Meeusen, W., & Van den Broeck, J. (1977). Efficiency estimation from cobb-douglas production functions with composed error. *International Economics Review, 18*(2), 435–444.

Meister, A. (2006). Density estimation with normal measurement error with unknown variance. *Statistica Sinica, 16*, 195–211.

Reifschneider, D., & Stevenson, R. (1991). Systematic departures from the frontier: A framework for the analysis of inefficiency. *International Economic Review, 32*, 715–723.

Richardson, W. H. (1972). Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America, 62*(1), 55–59.

Schwarz, M., Van Bellegem, S., & Florens, J. (2010). *Nonparametric frontier estimation from noisy data*. Springer-Verlag Berlin and Heidelberg GmbH & Co. K.

Stefanski, L. A., & Carroll, R. J. (1990). Deconvolving kernel density estimators. *Statistics, 21*, 169–184.

Wang, Y., Wang, S., Dang, C., & Ge, W. (2004). Nonparametric quantile frontier estimation under shape restriction. *European Journal of Operational Research*, 671–678.

Wu, H. S. (1990). Minimum entropy deconvolution for restoration of blurred two-tone images. *Electronics Letters, 26*(15), 1183–1184.

Publication II

Xiaofeng Dai. Non-parametric efficiency estimation in a panel setting: corrected Richardson-Lucy blind deconvolution. ***Proceedings of 2015 International Conference on Management Engineering and Information Technology Application*****, Hong Kong, China, April 19-20, 2015.**

© 2015 WIT Press Publications.
Reprinted with permission.

Non-parametric efficiency estimation in a panel setting: corrected Richardson-Lucy blind deconvolution

X. F. Dai

School of Biotechnology, JiangNan University, China
School of Business, Aalto University, Finland

Abstract

This paper presents a corrected Richardson-Lucy blind deconvolution method (cRLb) to decompose inefficiencies from the composite error coming from frontier estimation in a two-stage inefficiency estimation procedure. Simulations consisting of 19 scenarios show that cRLb is a non-parametric estimator that could obtain firm-specific inefficiency estimates, does not require pre-specification of the inefficiency distribution, is robust to data noise and heteroscedasticity, and generates matrix-form outputs given panel data.
Keywords: data envelopment analysis, corrected Richardson-Lucy blind deconvolution, efficiency estimation, nonparametric

1 Introduction

Productive efficiency analysis has two major branches, i.e., deterministic nonparametric approaches represented by DEA (data envelopment analysis) [1, 2] and stochastic parametric methods such as SFA (stochastic frontier analysis) [3, 4]. StoNED (stochastic semi-nonparametric envelopment of data) [5] builds the bridge between DEA and SFA and attracts increasing interest due to its semi-nonparametric and stochastic properties. A two-stage strategy is conventionally used for inefficiency identification when stochastic models are considered, where the frontier is estimated in the first stage using parametric or nonparametric regression techniques such as Modified Ordinary Least Squares (MOLS) [3] and Convex Nonparametric Least Squares (CNLS) [5] and the inefficiency is decomposed from the residuals (from the first stage) in the second stage using techniques such as method of moments (MM) in a cross-sectional setting or fixed effect approach [6] (abbreviated as SS here) in a panel setting. However, techniques such as MM and pseudolikelihood estimation require prior knowledge on inefficiency distribution and the fixed effect approach requires panel data to average out the random noise over time.

Richardson-Lucy blind deconvolution algorithm (RLb), originally developed for image deblurring, could be applied for inefficiency estimation [7]. The inefficiency term could be accurately decomposed from composite errors without the prior assumption on the distribution of inefficiencies u. Thus, RLb is exempted from issues such as the wrong skewness problem and is robust to the data noise level and heteroscedasticity [7]. However, estimates from RLb are upward biased by the expected inefficiency μ which is the same across firms and estimable using the method proposed by Hall and Simar [8, 9] from CNLS residuals in a three-stage strategy in the cross-section setting [7]. In the case of panel data, two models exist, i.e., the inefficiency is considered time-varying and time-invariant. If the inefficiencies are time-invariant, the three-stage strategy is still applicable where each time period is treated as an independent cross-section. However, when inefficiencies vary with time [10], such a method becomes insufficient. For this, we propose a corrected RLb algorithm (named cRLb) to remove μ from the inefficiency estimates. The cRLb method relies on a two-stage framework and fits both time-varying and time-invariant panel data. By simulating 19 scenarios including those from [3], we show that cRLb could effectively remove the the expected inefficiency when $\mu \neq 0$, and inherits all good properties of RLb such as no distributional assumption on the inefficiency and robustness to data noise and heteroscedasticity. Moreover, as an approach applied in the second step of the two-stage inefficiency estimation method, cRLb is easily coupled with any frontier estimation method.

The rest of the paper is organised as follows. The two-stage estimation strategy where cRLb is used for inefficiency decomposition in the panel setting and the fixed effects approach where the cRLb is evaluated against are described in the 'Method' section (more detailed information previously reported in [7, 11]). The 'Monte Carlo Simulation' section describes the data generation process of each tested scenario followed by

the results and discussion. The paper concludes by highlighting the novelties and contributions, summarizing the main results, and pointing out the future directions in the 'Conclusion' section.

2 Methods

2.1 Two-stage efficiency estimation

In a panel setting, considering the cost frontier model, $c_{i,t} = C(y_{i,t}) \exp(\varepsilon_{i,t})$, the cost $c_{i,t}$ of firm i at time t is computed from the cost function $C(y_{i,t})$, which is defined as the minimum cost of providing service $y_{i,t}$. The cost function C is a non-negative and non-decreasing function of the output $y_{i,t}$. The residual $\varepsilon_{i,t}$ could be decomposed into inefficiency $u_{i,t}$ and random noise $v_{i,t}$, i.e., $\varepsilon_{i,t} = v_{i,t} - u_{i,t}$ where $u_{i,t} \geq 0$. Mathematically, the problem is expressed as $\varepsilon_{i,t} = u_{i,t} + v_{i,t}$, where $v_{i,t}$ represents the noise contaminating and is independent of $u_{i,t}$. Note that in cRLb, no assumption is made for any of these parameters.

We consider a two-stage strategy for efficiency estimation. Both parametric and non-parametric models could be used for frontier estimation in the first step.

- Stage 1: Estimate the shape of cost function C by CNLS regression and obtain the residual $\hat{\varepsilon}_{i,t}$, where the model is defined as $c_{i,t} = C(y_{i,t}) \exp(\varepsilon_{i,t})$.
- Stage 2: Estimate the inefficiency $\hat{u}_{i,t}$ from the CNLS residuals obtained, using the fixed effects approach suggested by [6], i.e., the SS approach.

2.1.1 Stage 1: CNLS regression

The first step can be analytically represented by (1),

$$\begin{aligned} \min_{\alpha,\beta,\gamma,\hat{\varepsilon}} \sum_{i=1}^{I}\sum_{t=1}^{T} \hat{\varepsilon}_{i,t}^2 \qquad & \text{subject to} \\ \ln c_{i,t} = \ln(\alpha_{i,t} + \beta_{i,t}^{'} y_{i,t}) + \gamma^{'} Z_t + \hat{\varepsilon}_{i,t} \qquad & i = 1,\ldots,I; t = 1,\ldots,T \\ \alpha_{i,t} + \beta_{i,t}^{'} y_{i,t} \geq \alpha_{j,w} + \beta_{j,w}^{'} y_{i,t} \qquad & i,j = 1,\ldots,I; t,w = 1,\ldots,T \\ \beta_{i,t} \geq 0 \qquad & i = 1,\ldots,I; t = 1,\ldots,T \end{aligned} \tag{1}$$

where a contextual variable $\boldsymbol{Z}$ is used to model the inter-temporal relationship within the panel data where information on each decision making unit (DMU) i at each time point t was recorded [11]; and $\alpha_{i,t} = 0$ for all $i = 1 \ldots I$ and $t = 1 \ldots T$, under the assumption of constant return to scale (CRS). Note that α and β characterise tangent hyperplanes of the cost frontier, which provide a consistent estimator of $E(c|y)$ and are specific to each unit and time period. In particular, β can be interpreted as the marginal costs of the output y. γ is the coefficient of the contextual variable $\boldsymbol{Z}$, showing its weight. The set of inequality constraints is referred to as Afriat inequalities.

2.1.2 Stage 2: Corrected Richardson-Lucy blind deconvolution method

The cRLb method, the corrected blind form of the Richardson-Lucy algorithm applied in efficiency decomposition, is used in efficiency estimation provided with the residuals from the frontier estimation.

Let's first briefly go over the RL algorithm in the context of image recovery where it is originally developed for. According to [12], given the blurred image B and the clear image I, the intensity I_p at the pixel location p is computed from the pixel intensities B_q by $P(I_p) = \sum_q P(I_p|B_q)P(B_q)$ where $P(I_p)$ can be identified as the distribution of I_p and so forth. Expanding $P(I_p|B_q)$ by Bayes's rule, $P(I_p) = \sum_q \frac{P(B_q|I_p)P(I_p)}{\sum_z P(B_q|I_z)P(I_z)} P(B_q)$. The best of a bad situation is used to break the dependency of $P(I_p)$ on both sides, where the current estimation of $P(I_p)$ is used to approximate $P(I_p|B_q)$. Thus,

$$P^{j+1}(I_p) = \sum_q \frac{P(B_q|I_p)P^j(I_p)}{\sum_z P(B_q|I_z)P^j(I_z)} P(B_q) = P^j(I_p) \sum_q P(B_q|I_p) \frac{P(B_q)}{\sum_z P(B_q|I_z)P^j(I_z)}, \tag{2}$$

where j is the RL iteration index.

Considering $B^{'} = \sum_z P(B_q|I_z)P^j(I_z)$ as the predicted blurry image according to the current estimation of clear image I^j (a more workable notation for $P^j(I_z)$), define $P(B_q|I_z) = PSF(q,z)$, and use $E_q = \frac{B_q}{B_q^{'}}$ to denote the residual errors between the real and predicted blurry image, we obtain $\sum_q P(B_q|I_p)\frac{P(B_q)}{\sum_z P(B_q|I_z)P^j(I_z)} = \sum_q P(B_q|I_p)E_q^j$. If the isoplanatic condition holds, i.e., PSF is spatially invariant or $PSF(q,z)$ is the same for all q, then $B^{'} = \sum_z P(B_q|I_z)P^j(I_z) = I^j \otimes PSF$, and $\sum_q P(B_q|I_p)E_q$ becomes $PSF \star E_q$ where $\star$ and $\otimes$ are the correlation and convolution operators, respectively (note that the summation index in the generation of predicted blurry image, $B^{'}$, is z and for the integration of errors, E, is q). Hence, (2) becomes $I^{j+1} = I^j \times PSF \star \frac{B}{I^j \otimes PSF} = I^j \times PSF \star E^j$ where $E^j = \frac{B}{I^j \otimes PSF}$. Worth noting that the isoplanatic condition implies a symmetric assumption of the efficiency distribution which is usually asymmetric. Although the blind form of the RL algorithm is demonstrated to provide more precise inefficiency estimates than MM regarding the rankings [7], this assumption generally introduces an industry-wise bias which once adjusted could provide firm-specific inefficiency estimates.

The equations of the RL algorithm in the panel setting are analogous to those in the cross-sectional setting [7]. By identifying the inefficiency u as the clear image I, the residual ε as the blurry image B, and the noise v as the PSF, the iterative RL algorithm could be reformed as $u_{i,t}^{j+1} = u_{i,t}^j \times v \star \frac{\hat{\varepsilon}_{i,t}}{u_{i,t}^j \otimes v_{i,t}}$.

In the blind form of the RL algorithm, PSF (i.e., v here) is unknown and is iteratively estimated together with u. Let m be the index of the blind iteration, v_m is computed using (4) assuming that the object is known from the $(m-1)^{th}$ blind iteration

$$u_{i,t,m}^{j+1} = u_{i,t,m}^j \times v_m \star \frac{\hat{\varepsilon}_{i,t}}{u_{i,t,m}^j \otimes v_{i,t,m}}, \tag{3}$$

$$v_{i,t,m}^{j+1} = v_{i,t,m}^j \times u_{m-1} \star \frac{\hat{\varepsilon}_{i,t}}{u_{i,t,m-1} \otimes v_{i,t,m}^j}. \tag{4}$$

The RLb algorithm minimises the difference between the original and predicted degraded signals, i.e., $\arg\min_j(\varepsilon_{i,t} - \hat{\varepsilon}_{i,t})$, each firm at each time point with convergence proven in [13, 14]. However, this does not guarantee that the algorithm could find the global minimal if the cost function is not convex. In inefficiency estimation, the frontier is either parametrically determined or non-parametrically constrained to be convex, resulting in a convex function $\varepsilon_{i,t}$. Thus, under this context, RLb is guaranteed to find the global optimal at each data point. An initial guess of v_0^0 is required to start the algorithm, which is specified as 1's with the same dimension as the residuals here. The RLb algorithm requires non-negative inputs. For which, we add a large enough positive constant M to the inputs and subtract the deconvoluted M afterwards. The results are independent of the choice of M as tested by simulations.

The corrected form of RLb, i.e., cRLb, adjusts the inefficiency estimates from RLb by the difference between the average of the inefficiency estimates and that of the residuals over all DMUs in the cross-section case. It mathematically holds because

$$\hat{u}_i^* = \hat{u}_i - \overline{u} + \overline{u} = \hat{u}_i - \overline{u} + \overline{u} + \overline{v} = \hat{u}_i - (\overline{u} - \overline{\varepsilon}) \tag{5}$$

where $\hat{u}^*$ is the corrected estimate, $\overline{u}$ and $\overline{\varepsilon}$ are the averages of u and ε, respectively, across all DMUs, and $\overline{u} - \overline{\varepsilon} > 0$ is the bias corrected by cRLb, assuming $E(v) = 0$. In [7], this bias is corrected by Hall and Simar method [8] in the cross-sectional setting.

The cRLb method treats panel data naturally as it is and outputs the results with the same dimension. Thus, in the panel case, (5) becomes $\hat{u}_{i,t}^* = \hat{u}_{i,t} - (\frac{\sum_{t=1}^T \overline{u}_t}{T} - \frac{\sum_{t=1}^T \overline{\varepsilon}_t}{T})$, and is equivalent to $\hat{u}_{i,t}^* = \hat{u}_{i,t} - (\overline{u}_t - \overline{\varepsilon}_t)$ when u is time-invariant, where T represents the number of time points in the panel data. In this study, with a 2-dimentional matrix as the input, we obtain the efficiencies for each branch at each time point.

2.2 Fixed effects approach

The performance of cRLb is compared with that of SS [6] for inefficiency identification in the panel setting. In SS, the periodic performance of the sales unit $d_{i,t}$ is modelled as $\hat{d}_{i,t} = \exp(-\hat{\varepsilon}_{i,t})$. The average performance of unit i is obtained by $\hat{\overline{d}}_i = \sum_t^T \frac{\exp(-\hat{\varepsilon}_{i,t})}{T}$ and the efficiency is computed as $\text{Eff}_i = \frac{\hat{\overline{d}}_i}{\max(\hat{\overline{d}}_i)}$. In cRLb, the

average performance of unit i is obtained by $\hat{\bar{u}}_i = \sum_t^T \frac{-\hat{\varepsilon}_{i,t}}{T}$ and the inefficiency is computed as $\text{Eff}_i = \frac{\hat{\bar{u}}_i}{\max(\hat{\bar{u}}_i)}$.

3 Monte Carlo Simulation

3.1 Data generation process

With a similar design as in [7], we conducted two sets of simulations to assess the performance of the cRLb method with a data size of 50×100 (50 time points and 100 firms) for each simulation. The first set of simulations is analogous to [3], with the scenarios designed for different signal to noise ratios ($\lambda = \frac{\sigma_u}{\sigma_v}$) and two different models ($y = \varepsilon$ and $y = 1 + \varepsilon$), and the performance of cRLb is compared with RLb and SS. The second set of simulations aims at assessing the robustness of cRLb to the distributional assumption and data heteroscedasticity as compared against SS. Four distributions in addition to the half normal distribution, including 'truncated normal', 'gamma', 'exponential ' and 'Poisson', are tested. The noise term is assumed to follow normal distribution, with zero mean and a variance of 0.3. The signal to noise ratio in the second simulation set is set to the middle value ($\lambda = 1.24$) of the first simulation set under all scenarios. Group-wise heterscedasticity is generated for each heterscedastic data. Particularly, four equally divided sub-populations are generated, with consecutive data points being grouped together in their generic order (i.e., the first 25 data points belong to subgroup 1, points 26 to 50 belong to subgroup 2, and so on). The cRLb method and SS are applied to each scenario, with 100 iterations (Table 1).

Scenario	Parameter setting	λ	Model	Distribution
H2-0	$\mu_u = 0, \sigma_u = 0.8261$	1.66	$y = \varepsilon$	half normal
H1-0	$\mu_u = 0, \sigma_u = 0.6171$	1.24	$y = \varepsilon$	half normal
H0-0	$\mu_u = 0, \sigma_u = 0.4131$	0.83	$y = \varepsilon$	half normal
cH2-0	$\mu_u = 0, \sigma_u = 0.8261$	1.66	$y = \varepsilon$	half normal
cH1-0	$\mu_u = 0, \sigma_u = 0.6171$	1.24	$y = \varepsilon$	half normal
cH0-0	$\mu_u = 0, \sigma_u = 0.4131$	0.83	$y = \varepsilon$	half normal
cH2-1	$\mu_u = 0, \sigma_u = 0.8261$	1.66	$y = 1 + \varepsilon$	half normal
cH1-1	$\mu_u = 0, \sigma_u = 0.6171$	1.24	$y = 1 + \varepsilon$	half normal
cH0-1	$\mu_u = 0, \sigma_u = 0.4131$	0.83	$y = 1 + \varepsilon$	half normal
cH1-0	$\mu_u = 0, \sigma_u = 0.6171$	1.24	$y = \varepsilon$	half normal
cH1+0	$\mu_u = [0, 0, 0, 0], \sigma_u = [0.4131, 0.5508, 0.6884, 0.8261]$	1.24	$y = \varepsilon$	half normal
cT1-0	$\mu_u = 1, \sigma_u = 0.3882$	1.24	$y = \varepsilon$	truncated normal
cT1+0	$\mu_u = [0.8, 1.2, 0.5, 1.5], \sigma_u = [0.2598, 0.3465, 0.4331, 0.5197]$	1.24	$y = \varepsilon$	truncated normal
cG1-0	$\theta_u = 1, k_u = 0.1384$	1.24	$y = \varepsilon$	gamma
cG1+0	$\theta_u = [1, 1, 1, 1], k_u = [0.0620, 0.1102, 0.1722, 0.2480]$	1.24	$y = \varepsilon$	gamma
cE1-0	$\mu_u = 0.3720$	1.24	$y = \varepsilon$	exponential
cE1+0	$\mu_u = [0.2490, 0.3320, 0.4150, 0.4980]$	1.24	$y = \varepsilon$	exponential
cP1-0	$\mu_u = 0.1384$	1.24	$y = \varepsilon$	Poisson
cP1+0	$\mu_u = [0.0620, 0.1102, 0.1722, 0.2480]$	1.24	$y = \varepsilon$	Poisson
All	$\mu_v = 0, \sigma_v = 0.3$, data size = 50×100, iterations = 100			

Table 1: Parameter setting in the simulations. In gamma distribution, $\mu_u = \theta k, \sigma_u = \theta\sqrt{k}$. In exponential distribution, $\mu_u = \sigma_u = \theta\sqrt{k}$. In Poission distribution, $\sigma_u = \sqrt{\mu_u}$. Group-wise heteroskedasticity (four subgroups) is generated for 'cH1+0', 'cT1+0', 'cG1+0', 'cE1+0', 'cP1+0' (consecutive data points are grouped together in their generic order, i.e., data points of the first 25 columns (the first 50×25 data points) belong to subgroup 1 and so on); the elements in the square brackets are the corresponding parameters in each subgroup; and if more than one parameter is needed for a particular distribution, the elements are ordered in the same way in the brackets for different parameters.

Each scenario is given a four-digit name, with the scenarios tested using cRLb having a character 'c' in front. The four-digit name captures four aspects to be tested. Specifically, the first letter is the initial of the inefficiency distribution, i.e., 'H', 'T', 'G', 'E', 'P' are short for the half normal, truncated normal, gamma, exponential and Poisson distribution, respectively. The second digit shows the signal to noise ratio, which is represented by λ and defined as $\lambda = \frac{\sigma_u}{\sigma_v}$, with 2, 1, 0 representing the high, moderate and low levels, respectively. Here, this statistic is taken from [3], i.e., 2 is equivalent to $\lambda = 1.66$, 1 means $\lambda = 1.24$, and 0 is short for $\lambda = 0.83$. The third character indicates whether the data is heteroscedastic, where '-' and '+' each means without and with heteroscedasticity. The last digit shows the constant a in the model $y = a + \varepsilon$, i.e., $a = 0$ is associated with $y = \varepsilon$ and $a = 1$ is equivalent to $y = 1 + \varepsilon$.

3.2 Performance measures

Mean squared errors (MSE) are used to measure the performance of the algorithms in simulations, which are defined as $\text{MSE}_{\mu_u} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}_{u_i} - \mu_{u_i})^2, \text{MSE}_{\sigma_u} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}_{u_i} - \sigma_{u_i})^2, \text{MSE}_{\lambda} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\lambda}_i - \lambda_i)^2$, where μ_{u_i} (mean of u for DMU i), σ_{u_i} (standard deviation of u for DMU i) and $\lambda_i = \frac{\sigma_{u_i}}{\sigma_{v_i}}$ (signal to noise ratio) are estimated over 100 iterations for each simulated scenario, and $N = 100$.

3.3 Results and discussion

The statistics of the two sets of simulation results are summarised in Tables 2 and 3. The estimated and true inefficiencies are plotted against each other in Fig. 1 and Fig. 2.

Type	Stat	Met	H2-0	H1-0	H0-0	cH2-0	cH1-0	cH0-0	cH2-1	cH1-1	cH0-1
TRUE	μ_u		0.6590	0.4927	0.3300	0.6590	0.4927	0.3300	0.6590	0.4927	0.3300
EST	μ_u	cRLb	0.7904	0.6161	0.4280	0.6589	0.4927	0.3299	1.6589	1.4927	1.3299
MSE	μ_u	cRLb	0.0173	0.0152	0.0096	1.8E-5	1.7E-5	1.6E-5	0.9999	0.9999	0.9999
MSE	μ_u	SS	0.0090	0.0612	0.1378	0.0090	0.0612	0.1378	0.0510	0.1624	0.3293
TRUE	σ_u		0.4935	0.3690	0.2470	0.4935	0.3690	0.2470	0.4935	0.3690	0.2470
EST	σ_u	cRLb	0.0789	0.0645	0.0503	0.0787	0.0641	0.0493	0.0811	0.0676	0.0550
MSE	σ_u	cRLb	0.1720	0.0928	0.0387	0.1721	0.0930	0.0391	0.1701	0.0909	0.0369
MSE	σ_u	SS	0.1571	0.0692	0.0158	0.1571	0.0692	0.01580	0.2000	0.1061	0.0428
TRUE	λ		24.0452	17.9775	12.0332	24.0452	17.9775	12.0332	24.0452	17.9775	12.0332
EST	λ	cRLb	3.8414	3.1388	2.4468	3.8335	3.1212	2.4011	3.9518	3.2923	2.6769
MSE	λ	cRLb	412.4271	222.4429	92.8938	412.7391	222.9768	93.7841	407.9293	217.8808	88.4964
MSE	λ	SS	376.7731	165.8115	38.0478	376.7731	165.8115	38.0478	479.5823	254.1508	102.7014

Table 2: Results of simulation set 1. Scenarios with initial 'c' are tested using cRLb, and those without are run using RLb. 'H2-0', 'cH2-0': $y = \varepsilon$, $\lambda = 1.66$; 'H1-0', 'cH1-0': $y = \varepsilon$, $\lambda = 1.24$; 'H0-0', 'cH0-0': $y = \varepsilon$, $\lambda = 0.83$; 'cH2-1': $y = 1 + \varepsilon$, $\lambda = 1.66$; 'cH1-1': $y = 1 + \varepsilon$, $\lambda = 1.24$; 'cH0-1': $y = 1 + \varepsilon$, $\lambda = 0.83$. $\mu_v = 0$ and $\sigma_v = 0.3$ are used for data generation for all simulations. 100 simulations are run for each scenario. 'TRUE', 'EST', 'MSE' are the true value, estimated value, minimum standard error. 'Stat', 'Met' are short for statistics and method, respectively. Statistics are rounded to 4 digits.

Type	Stat	Met	cH1-0	cH1+0	cT1-0	cT1+0	cG1-0	cG1+0	cE1-0	cE1+0	cP1-0	cP1+0
TRUE	μ_u		0.4926	0.6574	1.0006	1.0308	0.1378	0.2551	0.3722	0.4965	0.1387	0.2548
EST	μ_u	cRLb	0.4928	0.6576	1.0007	1.0310	0.1379	0.2552	0.3723	0.4967	0.1388	0.2550
MSE	μ_u	cRLb	1.4E-5	1.6E-5	1.6E-5	4.7E-5	1.3E-5	1.9E-5	1.5E-5	1.5E-5	1.2E-5	1.7E-5
MSE	μ_u	SS	0.0621	0.0001	0.0234	0.2001	0.0809	0.0375	0.0919	0.0125	0.0842	0.0381
TRUE	σ_u		0.3696	0.4922	0.383	0.4756	0.3507	0.479	0.3667	0.4898	0.3668	0.4902
EST	σ_u	cRLb	0.0648	0.0807	0.0686	0.0786	0.0637	0.0792	0.0640	0.0803	0.063	0.078
MSE	σ_u	cRLb	0.0929	0.1759	0.0989	0.1648	0.0826	0.1670	0.0917	0.1741	0.0924	0.1765
MSE	σ_u	SS	0.0688	0.1726	0.1033	0.1893	0.0198	0.1167	0.0578	0.1540	0.0232	0.1240
TRUE	λ		17.9500	23.9363	18.6034	23.1282	17.0303	23.2879	17.8088	23.8133	17.8153	23.8330
EST	λ	cRLb	3.1454	3.9310	3.3314	3.8245	3.0957	3.8479	3.1071	3.9060	3.0633	3.7952
MSE	λ	cRLb	221.7659	421.9148	236.0365	395.7472	196.6467	401.0898	218.5740	417.3291	220.0686	423.2477
MSE	λ	SS	164.3076	414.3089	246.4501	454.8965	46.9563	280.3685	137.7277	369.2917	55.0114	297.2052

Table 3: Results of simulation set 2. 'cH1-0': half normal, homoscedastic; 'cH1+0': half normal, heteroscedastic; 'cT1-0': truncated normal, homoscedastic; 'cT1+0': truncated normal, heteroscedastic; 'cG1-0': gamma, homoscedastic; 'cG1+0': gamma, heteroscedastic; 'cE1-0': exponential, homoscedastic; 'cE1+0': exponential, heteroscedastic; 'cP1-0': Poisson, homoscedastic; 'cP1+0': Poisson, heteroscedastic. $y = \varepsilon$, $\lambda = 1.24$, $\mu_v = 0$, $\sigma_v = 0.3$ are used for data generation for all simulations, and 100 simulations are run for each scenario. 'TRUE', 'EST', 'MSE' are the true value, estimated value, minimum standard error. 'Stat' and 'Met' are short for statistics and method, respectively. Statistics are rounded to 4 digits.

3.3.1 cRLb removes μ from the inefficiency estimates

As illustrated by Fig. 1 'H2-0' to 'cH0-0', cRLb effectively removes the expected inefficiency μ regardless of the signal to noise ratio (λ). This property applies to data with different distributions of the inefficiency u and tolerates data heteroscedasticity (Fig. 1 'cH1-0' and Fig. 2). Statistically, the MSE of μ_u approximates 0 when cRLb is used ('cH2-0', 'cH1-0', 'cH0-0' of Table 2, and all scenarios in Table 3), whereas the MSE of σ_u and λ stay unchanged compared with the corresponding results using RLb ('H2-0', 'H1-0', 'H0-0' of Table 2), suggesting the accuracy of cRLb in correcting the bias introduced by RLb. On the other hand, cRLb does not correct the error coming from frontier estimation. This is illustrated by Fig. 1 'cH2-1' to 'cH0-1', where the estimates are upward shifted by the constant a in $y = a + e$ ($a = 1$ here). This is also statistically reflected in Table 2 where the MSEs of μ_u are almost 1 in 'cH2-1', 'cH1-1' and 'cH0-1'. As a benchmark for evaluating the performance of cRLb, SS always overestimates the inefficiency, and the bias increases with the decrease of λ.

3.3.2 cRLb inherits the advantages of RLb

RLb is shown to be robust to distribution density, skewness and kurtosis of u, as well as data noise and heterscedasticity [7]. These nice properties are inherited by cRLb and applicable to panel data. Except for the wrong skewness and low kurtosis problems (discussed in [7]) which do not exist using SS either, the other features inherited from RLb make cRLb outperforming SS in various aspects as described below.

As seen from Fig. 1 'cH1-0' and Fig. 2, cRLb is insensitive to the distribution assumption and always produces the actual firm-specific estimates, whereas the bias introduced by SS is highly affected by the distributional assumption of u. In most cases (gamma, exponential and Poisson distributions), SS overestimates

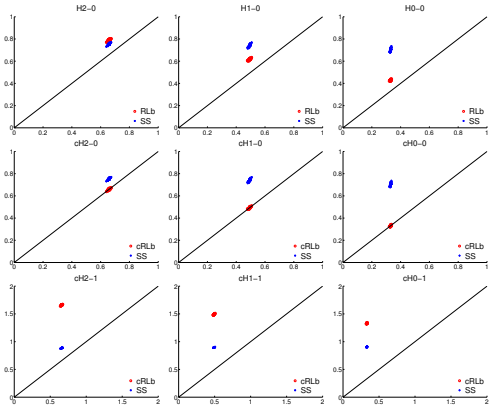


Figure 1: Simulation set 1 (half normal, homoscedastic). 'H2-0': RLb, $\lambda = 1.66$, $y = \varepsilon$; 'H1-0': RLb, $\lambda = 1.24$, $y = \varepsilon$; 'H0-0': RLb, $\lambda = 0.83$, $y = \varepsilon$; 'cH2-0': cRLb, $\lambda = 1.66$, $y = \varepsilon$; 'cH1-0': cRLb, $\lambda = 1.24$, $y = \varepsilon$; 'cH0-0': cRLb, $\lambda = 0.83$, $y = \varepsilon$; 'cH2-1': cRLb, $\lambda = 1.66$, $y = 1 + \varepsilon$; 'cH1-1': cRLb, $\lambda = 1.24$, $y = 1 + \varepsilon$; 'cH0-1': cRLb, $\lambda = 0.83$, $y = 1 + \varepsilon$.

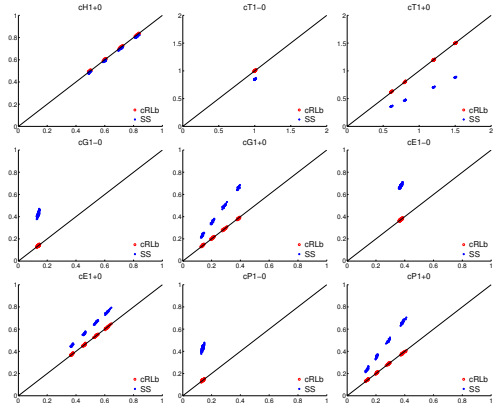


Figure 2: Simulation set 2 ($\lambda = 1.24$, $y = \varepsilon$). 'cH1-0': half normal, homoscedastic (see Fig. 1); 'cH1+0': half normal, heteroscedastic; 'cT1-0': truncated normal, homoscedastic; 'cT1+0': truncated normal, heteroscedastic; 'cG1-0': gamma, homoscedastic; 'cG1+0': gamma, heteroscedastic; 'cE1-0': exponential, homoscedastic; 'cE1+0': exponential, heteroscedastic; 'cP1-0': Poisson, homoscedastic; 'cP1+0': Poisson, heteroscedastic.

u with the bias increasing with the standard deviation of the inefficiency σ_u (Fig. 2 'cG1-0' to 'cP1+0'). When u follows truncated normal distribution, SS underestimates u with the bias increasing with σ_u (Fig. 2 'cT1+0'). Worth noting that when the distribution of u is half normal and the data is heteroscedastic, SS almost performs as equally well as cRLb. This means that when u follows a particular distribution such as half normal, SS could produce firm-specific inefficiency estimates supplemented with the additional information stored in the panel data (e.g., the periodic information and heteroscedasticity). Considering the truncated normal as a special case of half normal with a shifted mean, it is suggested that the bias introduced by SS is affected by the expected value of u and works best under half normal distribution. While cRLb is unaffected by such factors and always produces firm-specific estimates.

Inherited from RLb, cRLb is robust to data noise and heteroscedasticity. As illustrated in Fig. 1 'cH2-0' to 'cH0-0', estimates from cRLb lie on the line $\hat{u} = u$ regardless of λ while those from SS lie above and diverge from the line with the decrease of λ. Pairwise comparisons between the subplots of Fig. 1 'cH1-0' and Fig. 2 show that cRLb is insensitive to data heteroscedasticity, whereas SS is not.

4 Conclusions

This study presents a two-stage non-parametric inefficiency estimator, cRLb, which can produce firm-specific inefficiencies in the panel setting. The performance is tested under 19 simulated scenarios. The cRLb removes the expected inefficiency by the difference between the average of residuals ε and that of inefficiencies u over time for a specific firm in the panel setting, which is achieved by methods such as the non-parametric kernel estimator proposed by Hall and Simar [8] when RLb is used in the cross-sectional setting [7]. In other words, cRLb could estimate inefficiencies in a two-step framework provided with the panel data, whereas RLb could be used in the cross-sectional setting using a three-step strategy. Monte Carlo simulations show that the expected inefficiency μ could be effectively removed and all good properties of RLb such as non-parametric modelling, independence of inefficiency distribution, robustness to data noise and heteroscedasticity, are inherited by cRLb. The cRLb method could be applied to empirical cases such as estimating the inefficiencies of banking branches under a homogeneous environment or more complex cases such as a more diverse sales network in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant number: 31471251) and the Fundamental Research Funds for the Central Universities (grant number: JUSRP11507). I would like to thank Prof. Timo Kuosmanen for his insightful advices and help.

References

[1] Farrell, M.J., The measurement of productive efficiency. *Journals of the Royal Statistical Society*, 120(3), pp. 253-290, 1957.

[2] Charnes, A., Cooper, W.W., & Rhodes, E., Measuring the inefficiency of decision making units. *European Journal of Operational Research*, 2(6), pp. 429-444, 1978.

[3] Aigner, D.J., Lovell, C.A.K. & Schmidt, P., Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6(1), pp. 21-37, 1977.

[4] Meeusen W. & van den Broeck, J., Efficiency estimation from Cobb-Douglas production function with composed error. *International Economic Review*, 18(2), pp. 435-444, 1977.

[5] Kuosmanen T. & Kortelainen, M., Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis* 38(1), pp. 11-28, 2012.

[6] Schmidt, P. & Sickles, R.C., Production frontiers and panel data. *Journal of Bussiness and Economic Statistics*. 2(4), pp. 367-374, 1984.

[7] Dai, X.F., Non-parametric efficiency estimation using Richardson-Lucy blind deconvolution. *European Journal of Operational Research*, 248(2016), pp. 731-739, 2016.

[8] Hall, P. & Simar, L., Estimating a changepoint, boundary, or frontier in the presence of observation error. *Journal of the American Statistical Association*, 97(458), pp. 523-534, 2002.

[9] Delaigle, A. & Gijbels, I., Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics & Data Analysis*, 45(2), pp. 249-267, 2004.

[10] Cornwell, C., Schmidt, P., & Sickles, R.C., Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics*, 46(1-2), pp. 185-200, 1990.

[11] Eskelinen J. & Kuosmanen, T., Intertemporal efficiency analysis of sales teams of a bank: Stochastic semi-nonparametric approach. *Journal of Banking and Finance*, 37(12), pp. 5163-5175, 2013.

[12] Richardson, W., Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*. 62(1), pp. 55-59, 1972.

[13] Lucy, L.B., An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79(6), pp. 745-754, 1974.

[14] Irani, M. & Peleg, S., Improving resolution by image registration. *Graphical Models and Image Processing* 53(3), pp. 231-239, 1991.

Publication III

Xiaofeng Dai, Timo Kuosmanen. Best-practice benchmarking using clustering methods: Application to energy regulation. *Omega*, 42 (1), 179-188, 2014.

© 2014 Elsevier Publications.
Reprinted with permission.

Omega 42 (2014) 179–188

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

Best-practice benchmarking using clustering methods: Application to energy regulation

Xiaofeng Dai, Timo Kuosmanen *

Aalto University School of Business, 00101 Helsinki, Finland

ARTICLE INFO

Article history:
Received 21 August 2012
Accepted 21 May 2013
Available online 31 May 2013

Keywords:
Benchmark regulation
Productive efficiency
Data envelopment analysis (DEA)
Stochastic non-smooth envelopment of data (StoNED)
Clustering
Electricity distribution

ABSTRACT

Data envelopment analysis (DEA) is widely used as a benchmarking tool for improving productive performance of decision making units (DMUs). The benchmarks produced by DEA are obtained as a side-product of computing efficiency scores. As a result, the benchmark units may differ from the evaluated DMU in terms of their input–output profiles and the scale size. Moreover, the DEA benchmarks may operate in a more favorable environment than the evaluated DMU. Further, DEA is sensitive to stochastic noise, which can affect the benchmarking exercise. In this paper we propose a new approach to benchmarking that combines the frontier estimation techniques with clustering methods. More specifically, we propose to apply some clustering methods to identify groups of DMUs that are similar in terms of their input–output profiles or other observed characteristics. We then rank DMUs in the descending order of efficiency within each cluster. The cluster-specific efficiency rankings enable the management to identify not only the most efficient benchmark, but also other peers that operate more efficiently within the same cluster. The proposed approach is flexible to combine any clustering method with any frontier estimation technique. The inputs of clustering and efficiency analysis are user-specified and can be multi-dimensional. We present a real world application to the regulation of electricity distribution networks in Finland, where the regulator uses the semi-nonparametric StoNED method (stochastic non-parametric envelopment of data). StoNED can be seen as a stochastic extension of DEA that takes the noise term explicitly into account. We find that the cluster-specific efficiency rankings provide more meaningful benchmarks than the conventional approach of using the intensity weights obtained as a side-product of efficiency analysis.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

The purpose of benchmarking is to help the management of a decision making unit (DMU) to improve performance and productivity. The process of the best practice benchmarking involves the identification of the best firms in an industry or a sector, comparison of the specific performance metrics or indicators (e.g., unit cost, productivity, or efficiency), and learning from the peers how the business processes could be improved. The benchmarking process can be repeated continuously to allow DMUs improve their practices over time.

Data Envelopment Analysis (DEA) [1,2] has been widely applied for efficiency estimation and benchmarking (see, e.g., Section 3.9 of [3], and the recent surveys of DEA applications [4,5]). Technically, DEA is mainly geared towards efficiency estimation, applying input–output weights that maximize the efficiency score of the evaluated DMU. The conventional benchmarks provided by DEA can be seen as a side-product of the envelopment problem where the frontier is constructed as a convex hull of the observed data points using the so-called intensity weights (reference DMUs that have strictly positive intensity weights are identified as benchmarks, see Section 3.9 of [3]), while the benchmarks are widely considered as an appealing feature of DEA, to our knowledge, there is little evidence about the usefulness of the intensity weights for benchmarking (let alone their optimality). In the recent DEA literature (see [3] for an excellent survey), it is well recognized that units identified as benchmarks can differ from the evaluated DMU in terms of the input profile (e.g., capital intensity) or the output structure (economies of specialization versus scope). Further, the benchmarks can operate at different scale sizes than the evaluated DMU, particularly when constant returns to scale (CRS) is assumed. Indeed, if the benchmarks are located far away

Abbreviation: NMM, normal mixture model; StoNED, stochastic non-smooth envelopment of data; DEA, data envelopment analysis; DMU, decision making unit; SFA, stochastic frontier analysis; CNLS, Non-parametric methods including convex non-parametric least squares; CRS, constant returns to scale; VRS, variable returns to scale.

* Corresponding author. Tel.: +358943131.
E-mail addresses: xiaofeng.dai@me.com (X. Dai), timo.kuosmanen@aalto.fi (T. Kuosmanen).

0305-0483/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.omega.2013.05.007

from the evaluated DMU in the input–output space, the benefits of the benchmarking exercise may be questionable.

The benchmark selection has attracted growing interest in the recent DEA literature: there is a growing stream of DEA studies on the identification of closest targets, axiomatic characterization of benchmarks, and the use of preference information and interactive procedures (see, e.g., Refs. [6–11] and Section 3.9 of [3] for further discussion). To our knowledge, however, these recent developments restrict to the deterministic DEA framework that assumes away noise. It is well recognized that DEA can be sensitive to random noise and heterogeneity of DMUs and their operating environments. In a stochastic environment, some DMUs may appear more efficient than others due to more favorable operational conditions or just pure luck (consider, e.g., external demand factors or weather conditions), while DEA can identify successful units, it may be difficult to transfer the success recipes to inefficient DMUs if the success is due to external conditions or just good fortune.

The motivation of this paper stems from a real-world application to the regulation of electricity distribution networks, which is one of the most significant application areas of DEA and efficiency analysis. Traditionally, regulators in many countries have applied DEA to estimate the efficient frontier to serve as the best practice benchmark in the regulatory framework. In the past decade, several countries have adopted stochastic frontier analysis (SFA [12,13]) models to complement DEA. The main advantage of SFA is that it models the random noise term explicitly in a probabilistic manner. However, the SFA imposes more restrictive parametric functional form assumptions than DEA. Recently, the Finnish regulator (Energiamarkkinavirasto EMV) replaced the conventional DEA and SFA by the new StoNED method (stochastic non-parametric envelopment of data [14,15]). The StoNED method combines the appealing features of both DEA and SFA, melding the axiomatic DEA-style non-parametric frontier with the probabilistic SFA-style treatment of noise. The StoNED method differs from the semi-parametric extensions of SFA in that it does not make any assumptions about the functional form or its smoothness (see [14] for a more detailed discussion). Rather, StoNED builds directly on the axioms of production theory (such as free disposability and convexity), similar to DEA. Compared to DEA, the StoNED method differs in its probabilistic treatment of inefficiency and noise, while the DEA frontier is typically spanned by a small number of influential observations, which makes it sensitive to outliers and noise, the StoNED method uses information of all observations in the data set to estimate the frontier. The StoNED method can also be applied to panel data (see [14]) and the observed heterogeneity of units and their operating environments can be explicitly modeled as an integral part of the estimation (see [16,17]).

Benchmarking forms an integral part of the frontier based regulatory regimes. As inefficient energy companies are required to reduce their total costs, it is necessary to indicate companies that provide comparable service in a similar environment with a lower cost. Of course, the conventional approach is to identify benchmarks based on the intensity weights, and this could be used equally well in DEA and StoNED. In the present application, however, many energy companies find the conventional benchmarks inappropriate. Finland is a sparsely populated country with a relatively large land area covered by forest and lakes. As a result, the Finnish electricity distribution sector consists of a very heterogenous group of firms. Some firms operate in larger cities such as Helsinki, where underground cables form a large proportion of the electricity grid. A majority of firms operates in rural areas, using overhead cables. There are also some small firms which are specialized to supply power to industrial users. The main problem with the conventional DEA benchmarks is that often urban network companies are identified as benchmarks for rural network firms, and vice versa. It is necessary to take the heterogeneity of firms explicitly into account in the benchmarking procedure.

To identify more appropriate benchmarks, in this paper we propose a novel approach based on the clustering methods, which applies equally well to the conventional DEA and SFA as well as to the recently introduced StoNED method. The proposed approach can be briefly described as follows. We apply a certain clustering method to identify a number of mutually exclusive groups from the original input–output data, or from the input–output vectors that are first projected to the estimated frontier. In each cluster, we rank the DMUs in the descending order of efficiency. These cluster-specific rankings allow managers to identify not only the best performing DMUs within each group, but also a range of DMUs that performs better within the same cluster. The full range of efficiency scores within a cluster can provide managerial insights into why some DMUs are more efficient than others within the same cluster, and help the managers to identify the most appropriate benchmarks, both in the short run and long run.

We must recognize that clustering methods have been used in the context of efficiency analysis before. For example, the latent class SFA models identify groups of DMUs which are interpreted to operate with different technologies (see, e.g., [18]). O'Donnell et al. suggested using clustering methods to identify latent classes in the context of meta-frontier estimation [19]. In the DEA literature, Po et al. proposed to apply DEA as a clustering technique [20]. Triantis et al. presented a two-stage strategy for efficiency performance analysis [21]. Fallah-Fini et al. proposed a bootstrapped non-parametric meta-frontier approach to measure the efficiency of highway maintenance contracting strategies [22]. To summarize, the previous studies that combine clustering approaches with efficiency analysis restrict to specific clustering method or to particular applications. To our knowledge, this paper is the first one to apply clustering methods specifically for benchmarking purposes.

The general approach to benchmarking proposed in this paper is highly flexible. It applies to any frontier estimation method, including DEA, SFA, and StoNED. Further, any appropriate clustering technique may be applied. Since there exists a large literature of clustering methods, we present a concise survey of methods, classified as hierarchical, partitioning, and model-based clustering methods. The approach is also flexible in terms of the clustering criteria. One can use the input–output variables, some functions thereof, or some other observed characteristics of the firm as input data to clustering. One can apply different techniques or combinations thereof to gain better understanding of which DMUs are similar to the evaluated unit, and which criteria can best characterize similarity. The choice of the criteria and the clustering method can be conducted interactively with the management to ensure the maximum relevance for the decision makers.

The rest of the paper is organized as 'theory', 'application' and 'conclusion'. In the next section we introduce the frontier production model, briefly review the DEA and StoNED approaches, summarize the widely used clustering methods, and elaborate our proposal for the benchmarking framework. Section 3 presents the real world application for the regulation of electricity distribution networks in Finland, and discusses some implementation issues. Finally, Section 4 concludes.

2. Theory

The proposed clustering based benchmarking framework incorporates frontier estimation and clustering methods into a unified flexible framework. As the two main steps in the framework for

benchmarking, any frontier estimation and clustering methods could be employed in principle. Thus, in this section, we first introduce the widely available techniques in each of the two steps, and formally present the framework in the end.

2.1. Frontier estimation methods

The field of productive efficiency analysis has been dominated by non-parametric DEA [1,2] and parametric SFA [12,13]. The appeal of DEA lies in its non-parametric nature (i.e., no functional form assumption for the frontier), whereas SFA appeals with its stochastic treatment of the deviations which is decomposed into a non-negative inefficiency term and a random noise term. The emergence of the StoNED framework (stochastic non-parametric envelopment of data [14]) bridges the gap between DEA and SFA. StoNED is a semi-parametric method that encompasses DEA and SFA as its special cases. Any of these methods can be easily incorporated into the proposed benchmarking framework.

Consider the standard multiple-input $\mathbf{x}_i$, single-output y_i, cross-sectional model,

$$\begin{aligned} y_i &= f(\mathbf{x}_i) + \varepsilon_i, \\ &= f(\mathbf{x}_i) - u_i + v_i, \quad \forall i = 1, \ldots, N. \end{aligned} \tag{1}$$

where $u_i > 0$ is an asymmetric inefficiency term and v_i is a stochastic noise term. DEA, SFA and StoNED are formulated depending on how the production function f and the random variables u and v are estimated. In the following we focus on DEA and StoNED.

2.1.1. DEA

DEA is deterministic in the sense that the stochastic noise term v is assumed away. Instead of assuming any particular form for the production function, DEA assumes that f satisfies certain regulatory axioms, i.e., monotonicity and concavity. The variable returns to scale (VRS) DEA estimator of f can be defined as [23,24]

$$f^{DEA}(\mathbf{x}) = \max_{\lambda \in \Re_+^N} \left\{ y | y = \sum_{h=1}^{N} \lambda_h y_h;\ \mathbf{x} \geq \sum_{h=1}^{N} \lambda_h \mathbf{x}_h;\ \sum_{h=1}^{N} \lambda_h = 1 \right\}, \tag{2}$$

and the efficiency estimate ε_i^{DEA} for DMU i can be obtained by substituting f in (1) by (2), as seen in (3) [24]

$$\begin{aligned} \varepsilon_i^{DEA} = \min_{\lambda, \varepsilon} \Bigg\{ & \varepsilon | y_i = \sum_{h=1}^{N} \lambda_h y_h + \varepsilon; \\ & \mathbf{x}_i \geq \sum_{h=1}^{N} \lambda_h \mathbf{x}_h;\ \sum_{h=1}^{N} \lambda_h = 1;\ \lambda_h \geq \mathbf{0}, \quad \forall h = 1, \ldots, N \Bigg\}, \end{aligned} \tag{3}$$

Problem (3) minimizes ε, which represents inefficiency: DMUs that yield $\varepsilon = 0$ are classified as efficient. Problem (3) also includes intensity weights λ, which are used for constructing convex combinations of the observed DMUs. Reference units that have a positive value of weight λ in the optimal solution to (3) are conventionally used as the benchmarks for the evaluated DMU. As noted in the introduction (Section 1), however, the input–output profiles of the evaluated DMU and the units identified as benchmarks can differ considerably. Further, the evaluated DMU might operate at different scale sizes than the benchmarks, especially if CRS is assumed. It can be argued that the further away the benchmarks are located from the evaluated DMU, the more difficult it will be to transfer the knowledge and practices of the benchmark units to the evaluated DMU. Aparicio et al. [25] recognized this problem, stating the following: *"The DEA models yield targets that are usually determined by the "furthest" efficient projection to the assessed DMU…. However, we believe, as many other authors, that the projected point on the efficient frontier obtained as such may not be a representative projection for the assessed DMU and that the distance to this efficient projection should be minimized so that the resulting targets are as much similar as possible to the inputs and outputs of the assessed DMU. The general argument behind this idea is that closer targets suggest directions of improvement for the inputs and outputs of the inefficient units that may lead them to the efficiency with less effort."* To address the problem, they proposed methods such as the Euclidean distance-based measure [26] to obtain the shortest path to the efficient frontier from the assessed DMU, while Aparicio et al. state their argument in the context of target setting, in our view, the same argument applies to benchmarking.

DEA is a deterministic method in the sense that it attributes all deviations from the frontier to inefficiency u, and hence assumes away the noise term v, while the presence of stochastic noise is not necessarily a problem if one is mainly interested in identifying the best-performing units in the sample, it does affect the benefits of benchmarking. If the efficiency differences are to a large extent driven by random factors that are beyond the control of the management, transferring the good practices becomes challenging. For example, it is possible that DMU *A* has better practices than DMU *B*, but due to random errors, DMU *B* is classified as DEA efficient whereas DMU *A* appears inefficient. Clearly, benchmarking involves risks in a noisy environment. We next consider the semi-nonparametric StoNED method, which takes the random noise term v explicitly into account.

2.1.2. StoNED

The StoNED method combines the non-parametric, piece-wise linear DEA-style frontier with the stochastic SFA-style treatment of inefficiency and noise. The assumptions of StoNED are milder than those required by DEA or SFA: both DEA and SFA can be obtained as constrained special cases of the more general StoNED-model (see [14]). The less restrictive assumptions directly imply that StoNED has a wider range of applicability: it is more robust to uncertainty concerning both the functional form of the frontier and the stochastic noise. The model is defined as (1), where f has no particular functional form but satisfies monotonicity and concavity. A two-stage strategy is used to estimate the deterministic part of the StoNED model in a non-parametric fashion [14]. In the first stage, the shape of the function f is estimated by CNLS regression, given that DEA can be interpreted as CNLS that is subject to the sign constraints on residuals. In the second stage, by imposing additional distributional assumptions, e.g., the asymmetric distribution for u_i with positive mean μ and finite variance σ_u^2, and a symmetric distribution for v_i with zero mean and constant finite variance σ_v^2, the variances are estimated based on the skewness of the CNLS residuals obtained from Stage 1 using the method of moments or pseudo likelihood techniques. The inefficiency u is then computed from the variance parameter estimates.

Specifically, the problem can be analytically represented by (4)–(7) [14],

$$\min_{\mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}} \sum_{i=1}^{n} \varepsilon_i^2 \quad \text{such that} \tag{4}$$

$$y_i = \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i + \varepsilon_i \tag{5}$$

$$\alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}_h' \mathbf{x}_i, \quad \forall h, i = 1, \ldots, n \tag{6}$$

$$\boldsymbol{\beta}_i \geq \mathbf{0}, \quad \forall i = 1, \ldots, n \tag{7}$$

where α_i and $\boldsymbol{\beta}_i$ are coefficients specific to observation i and v_i captures its random noise.

The inefficiency is then computed using the distribution of the CNLS residuals $\hat{\varepsilon}_i$ (note that $\varepsilon = v_i + u_i$). With the assumption that the inefficiency and noise follow half-normal and normal distribution,

respectively, the 2nd and 3rd central moments of the composite error distribution are

$$M_2 = \left[\frac{\pi-2}{\pi}\right]\sigma_u^2 + \sigma_v^2, \tag{8}$$

$$M_3 = -\left(\sqrt{\frac{2}{\pi}}\right)\left[\frac{4}{\pi}-1\right]\sigma_u^3, \tag{9}$$

which can be estimated using the CNLS residuals

$$\hat{M}_2 = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \bar{\varepsilon})^2/n, \tag{10}$$

$$\hat{M}_3 = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \bar{\varepsilon})^3/n. \tag{11}$$

Thus, the standard deviations of the inefficiency and error term are

$$\hat{\sigma}_u = \sqrt[3]{\frac{\hat{M}_3}{\left(\sqrt{\frac{2}{\pi}}\right)\left[\frac{4}{\pi}-1\right]}}, \tag{12}$$

$$\hat{\sigma}_v = \sqrt{\hat{M}_2 - \left[\frac{\pi-2}{\pi}\right]\hat{\sigma}_u^2}. \tag{13}$$

Since the conditional distribution of the inefficiency u_i given ε_i is a zero-truncated normal distribution with mean $\mu_\star = -\varepsilon_i\sigma_u^2/(\sigma_u^2+\sigma_v^2)$ and variance $\sigma_\star^2 = \sigma_u^2\sigma_v^2/(\sigma_u^2+\sigma_v^2)$, the conditional mean can be computed as

$$E(u_i|\varepsilon_i) = \mu_\star + \sigma_\star\left[\frac{\phi(-\mu_\star/\sigma_\star)}{1-\Phi(-\mu_\star/\sigma_\star)}\right], \tag{14}$$

where ϕ and Φ represent the standard normal density function and the standard normal cumulative distribution function, respectively.

It is worth to note that the frontier characterized by the CNLS problem (4)–(7) is not necessarily unique. Kuosmanen and Kortelainen [14] recognize this problem. The solution they propose is to use the lower bound of the set of piece-wise linear functions that solve the problem (4)–(7). Further, in Theorem 3.2 (of [14]) they formally show that lower bound is obtained by applying the standard DEA method to the projection points $(\mathbf{x}, \hat{y})$ of the CNLS problem. Thus, the intensity weights λ of the DEA problem used for characterizing the StoNED frontier could be used for benchmarking purposes in the same way as they are conventionally used in DEA. However, the arguments presented in Section 2.1.1 still apply: there is no guarantee that the benchmarks have a similar input–output structure or scale size as the evaluated DMU. Furthermore, since the observed DMUs are first projected to the StoNED frontier before DEA is applied, there is no guarantee that the DMUs identified as benchmarks are efficient. This observation motivates us to consider an alternative approach to benchmarking.

2.2. Clustering methods

Clustering is a common technique used in many fields including, e.g., bioinformatics, image analysis, pattern recognition and information retrieval. Specifically, functionally related genes can be grouped together to reveal novel pathways or new functions of certain genes; image boundaries can be easily located in image segmentation using clustering; as an unsupervised pattern recognition method, clustering can be used to identify patterns based on the similarities within the data; and information similar to the queries can be retrieved from documents by clustering. In the proposed benchmarking framework, DMUs are first clustered into groups from the original input–output data. The segmentation is dependent on the clustering algorithm, especially when the group boundaries are ambiguous. Thus, choosing the appropriate clustering algorithm is fundamental in obtaining meaningful benchmarking results. Below we review some widely used clustering algorithms.

2.2.1. Hierarchical methods

There are two types of hierarchical clustering algorithms, namely the agglomerative method and the divisive method, which recursively combines or splits a set of objects into bigger or smaller groups based on a certain criterion [27,28]. Commonly applied criteria include single linkage [28], complete linkage [28], average linkage [28], group average linkage [28] and Ward's linkage [29,28]. The group similarity is often scaled by distance, for which different measurements can be employed depending on the purpose and the characteristics of the firms. Among others, Euclidean distance [30], Mahalanobis distance [31], Manhattan distance [32], and Hamming distance [33] are most commonly seen. The formulations of the aforementioned clustering criteria and the distance measures are presented in Appendix A.

Hierarchical clustering is favored due to its simple yet intuitively reasonable principle. However, it requires expert domain knowledge to define the distance measurement for a particular problem. For example, Euclidean distance is suitable when the data is representable in vector space but should be avoided in high-dimensional text clustering [34]. Moreover, the number of clusters depends highly on the granularity chosen by the user, rendering the results subjective to the pre-assumptions [35]. Also, outliers, if exist, may distort the clustering results.

2.2.2. Partitioning methods

Partitioning methods are another class of heuristic methods besides hierarchical clustering. The principle is to iteratively reallocate data points across groups until no further improvement is obtainable [36,35]. K-means [36] is a typical and the most representative partitioning algorithm. It is based on the criterion that each object belongs to its closest group, where the group is represented by the mean of its objects. In particular, with a given g, the algorithm partitions N observations, $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$, into g groups ($\mathbf{G} = \{G_1, G_2, \ldots G_g\}$) by minimizing the total intra-cluster variance, i.e., $\text{argmin}_{\mathbf{G}} \sum_{i=1}^{g} \sum_{\mathbf{r}_w \in \mathbf{G}_i} (\mathbf{r}_w - \mu_i)^2$, where μ_i is the mean of $\mathbf{G}_i$.

It is seen from K-means that the number of clusters has to be prespecified or known. Also, the clustering results may be contaminated by outliers [35]. Successive efforts have been devoted to search their remedies which, however, mostly involve techniques out of the domain of partitioning methods. For example, X-means (extended from K-means) solves the problem of selecting the number of clusters via using model selection criteria [37].

Despite those disadvantages, partitioning methods are widely applied due to their simplicities. Many algorithms, such as fuzzy C-means [38], quality threshold clustering [39] and partitioning around medoids [40], also belong to this category. Specifically, 'fuzzy C-means' assigns each data point to each cluster with a certain probability [38], 'quality threshold' groups data points whose similarities are high enough together [39], and 'partitioning around medoids' minimizes a sum of dissimilarities and allows the user to choose the number of clusters through graphical display [40].

2.2.3. Model based clustering

Model based methods attempt to optimize the fitness between the data and the model where the data is assumed to be generated [41–44]. Model based methods can be further classified into finer groups, including finite mixture models [41], infinite mixture models [42], model based hierarchical clustering [43], and specialized model based partitioning clustering [44], among which finite model based methods are most widely applied.

In finite model based clustering, each observation $\mathbf{r}$ is drawn from finite mixture distributions with the prior probability π_i,

component-specific distribution f_i and parameters $\boldsymbol{\theta}_i$. The formula is given as

$$f(\mathbf{r};\Theta)=\sum_{i=1}^{g}\pi_i f_i(\mathbf{r};\boldsymbol{\theta}_i), \tag{15}$$

where $\Theta=\{(\pi_i,\boldsymbol{\theta}_i): i=1,\ldots,g\}$ is used to denote all unknown parameters, with the restriction that $0\le\pi_i\le 1$ for any i and $\sum_{i=1}^{g}\pi_i=1$. Note that g is the number of components in this model.

Expectation Maximization (EM) algorithm is normally used for the above model based clustering. The data log-likelihood can be written as

$$\log L(\Theta)=\sum_{j=1}^{N}\log\left(\sum_{i=1}^{g}\pi_i f_i(\mathbf{r}_j;\boldsymbol{\theta}_i)\right), \tag{16}$$

where $R=\{\mathbf{r}_j: j=1,\ldots,N\}$ and N is the total number of observations.

Since direct maximization of (16) is difficult, the problem can be casted in the framework of incomplete data. Define I_{ji} as the indicator of whether $\mathbf{r}_j$ comes from component i, the complete data log-likelihood becomes

$$\log L_c(\Theta)=\sum_{j=1}^{N}\sum_{i=1}^{g}I_{ji}\log(\pi_i f_i(\mathbf{r}_j;\boldsymbol{\theta}_i)). \tag{17}$$

At the mth iteration of the EM algorithm, the E step computes the expectation of the complete data log-likelihood which is denoted as Q

$$\begin{aligned}Q(\Theta;\Theta^{(m)})&=E_{\Theta^{(m)}}(\log L_c|R)\\&=\sum_{j=1}^{N}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log(\pi_i f_i(\boldsymbol{I}_j;\boldsymbol{\theta}_i)),\end{aligned} \tag{18}$$

and the M step updates the parameter estimates to maximize Q. The algorithm is iterated until convergence. Note that I's in (17) are replaced with τ's in (18), and the relationship between these two parameters is $\tau_{ji}=E[I_{ji}|\mathbf{r}_j,\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_g;\hat{\pi}_1,\ldots,\hat{\pi}_g]$. The set of parameter estimates $\{\hat{\boldsymbol{\theta}}_1,,\hat{\boldsymbol{\theta}}_g;\hat{\pi}_1,\ldots,\hat{\pi}_g\}$ is a maximizer of the expected log-likelihood for given τ_{ji}'s, and we can assign each $\mathbf{r}_j$ to its component based on $\{i_0|\tau_{ji_0}=\max_i\tau_{ji}\}$.

One advantage of mixture model based clustering is its automatic determination of the number of clusters. Commonly used model selection criteria can be roughly classified as likelihood based methods [45] and approximation based methods [46–51], where four approximation based model selection criteria are widely applied due to their computational efficiency, which are Akaike information criterion (AIC) [47,50], modified AIC (AIC3) [49,50], Bayesian information criterion (BIC) [48,51], and integrated classification likelihood BIC (ICL-BIC) [46].

2.3. Clustering framework for benchmarking

The proposed benchmarking framework combines clustering and productive efficiency analysis into a unified framework. It first clusters the DMUs into groups based on user-specified metrics, and then identifies relative or absolute benchmarks using productive efficiency analysis. We define the 'relative benchmark' as a DMU 'h' that achieves the highest efficiency in the group but falls below 100% efficiency; and the 'absolute benchmark' as a DMU 'h' which achieves at least 100% efficiency.

Mathematically, assume that N DMUs are clustered into g groups using a particular clustering algorithm, and there are N_{G_j} DMUs in cluster G_j. Let ζ_i denote the efficiency of DMU i ($i\in 1,\ldots,N_{G_j}$ in group G_j), and denote the frontier DMU(s) as h, the relative and absolute benchmarks are defined by (19) and (20).

$$\text{Relative benchmark}: h=\left\{i|\max_{i=1}^{N_{G_j}}\zeta_i\right\},\quad \max_{i=1}^{N_{G_j}}\zeta_i<1 \tag{19}$$

$$\text{Absolute benchmark}: h=\left\{i|\zeta_i\ge 1\right\},\quad \max_{i=1}^{N_{G_j}}\zeta_i\ge 1 \tag{20}$$

To qualify as absolute benchmark, the DMU must operate with 100% efficiency or higher. It is possible that there are multiple absolute benchmarks within the same cluster. It is also possible that all DMUs within a cluster are inefficient. In this case, we propose to indicate the DMU with the highest efficiency score within the cluster as a relative benchmark. The rationale of the relative benchmark is similar to that of the context dependent DEA discussed in [52].

The 'relative benchmark' is provided when none of the DMUs in a particular group has 100% efficiency due to, e.g., the cluster-wise operational inefficiencies, ensuring at least one reference for each inefficient DMU to benchmark against. The 'absolute benchmark' must have at least 100% efficiency, allowing multiple choices for a particular DMU in a given group.

In a step-by-step manner, the clustering based benchmarking framework runs as below:

- Step 1: efficiency estimation. Compute the efficiency score for the whole data set using a productive efficiency estimation method such as StoNED, DEA, and SFA.
- Step 2: clustering. Group the DMUs using a clustering algorithm such as mixture models and K-means. The data used for clustering can include any user specified metrics such as the inputs, outputs and efficiency scores from productive efficiency analysis, depending on the objectives and scope of the benchmarking application.
- Step 3: benchmarking. Find the absolute or relative benchmark (s) for DMUs of each group using the efficiency scores computed from the first step.

Alternatively, when the number of DMUs is sufficiently large, the procedure can be implemented in the reverse order:

- Step 1: clustering. Group the DMUs using a clustering algorithm such as mixture models and K-means. The inputs for clustering can be any user specified metrics depending on the aspects needed for benchmarking.
- Step 2: efficiency estimation. Compute the efficiency score within each group using a productive efficiency estimation method such as StoNED, DEA, and SFA.
- Step 3: benchmarking. Find the absolute benchmark(s) for DMUs of each group using the efficiency scores computed from the second step.

Note that in the second alternative, only the absolute benchmarks are produced.

In the application to energy regulation, to be described and discussed in the next section, we prefer to apply the combination of StoNED for frontier estimation and efficiency analysis and NMM for clustering. In other applications, one may prefer some other combination of methods. Indeed, we see the generality and flexibility of the proposed framework as major advantages of the proposed framework. The data used in clustering and efficiency analysis are user-specified, and can be multi-dimensional. The combination of methods applied for clustering and efficiency analysis can be chosen to meet the objectives and scope of the application. Finally, the order in which clustering and efficiency analysis are applied can be reversed, provided that the sample size

is sufficiently large such that it is meaningful to apply efficiency analysis separately in each cluster.

3. Application

3.1. Data and methods

3.1.1. Data

We applied the clustering based cluster-specific benchmarking framework to Finnish electricity distribution networks. The data consists of the six-year average over the period 2005–2010, which is available in the Energy Market Authority (EMA) website (www.emvi.fi). The cost frontier model has been adopted by EMA, where the total cost (x) is used as the single input, and three variables, i.e., 'Energy transmission' (GWh of 0.4 kV equivalents, y_1), 'Network length' (km, y_2), and 'Customer number' (y_3) are specified as the outputs ($\mathbf{y}$). Specifically, x includes the operational expenditure and half of the interruption cost, and the electricity transmission at different voltage levels is weighted according to the average transmission cost such that lower weight is assigned to high-voltage transmission than low-voltage transmission in y_1. To better control the DMUs' heterogeneity and their operating environment, the proportion of the underground cables in the total network length is used as a contextual variable (z). The descriptive statistics of the data are listed in Table 1. A more detailed description of the variables and the regulatory application is presented in [15]. The complete set of input and output data used in this application is provided as supplementary material to this article, available online at ⟨http://www.sciencedirect.com⟩.

3.1.2. Methods

StoNED and Normal Mixture Model (NMM) are used for productive efficiency analysis and group clustering, respectively, in this empirical study. The results are compared with those from DEA in Section 3.3.

In the efficiency estimation, the current regulatory model of EMV, i.e., the cost frontier model, is used. In Finland, underground cables are widely used in urban and suburban regions but not in rural areas, we thereby use a contextual variable z to capture the heterogeneity introduced by the proportion of underground cables. Following [15], we specify the cost frontier model as

$$x_i = C(\mathbf{y}_i)\exp(\delta z_i + u_i + v_i), \tag{21}$$

where C denotes the frontier cost function and δ characterizes the effect of underground cables z on a DMU's total cost. By taking the natural logarithms on both sides of (21), the parameters can be estimated by convex programming using the following equations,

$$\min_{\gamma,\boldsymbol{\beta},\delta,\boldsymbol{\varepsilon}} \sum_{i=1}^{n} \varepsilon_i^2 \quad \text{such that} \tag{22}$$

$$\ln x_i = \ln \gamma_i + \delta z_i + \varepsilon_i, \quad \forall i \tag{23}$$

$$\gamma_i = \boldsymbol{\beta}'_i \mathbf{y}_i, \quad \forall i \tag{24}$$

$$\gamma_i \geq \boldsymbol{\beta}_h' \mathbf{y}_i, \quad \forall h, i \tag{25}$$

$$\boldsymbol{\beta}_i \geq \mathbf{0}, \quad \forall i \tag{26}$$

where $\varepsilon_i = v_i + u_i$ and γ_i is the CNLS estimator of $E(x_i|\mathbf{y}_i)$.

In the present application, we take the stochastic noise term v explicitly into account in the estimation of the cost frontier C. However, we must stress that the inefficiency term u is considered as a random variable: its realization cannot be consistently estimated based on just a single observation. No consistent estimator of the inefficiency term u is available in the parametric SFA literature (in the cross-sectional setting considered here). Since in the present application the regulator is mainly interested in the frontier (to specify the efficient cost levels for each DMU), we measure efficiency as distance from the observed DMU to the frontier. Note that the data of observed DMUs is also likely to contain noise. As the purpose is to set targets from the cost frontier, we do not adjust the efficiency estimates for the noise in the observed data of units. That is, we measure efficiency as the ratio of the efficient cost and the actual cost, i.e., $C(\mathbf{y}_i)/x_i$. In this case, it is possible that these efficiency scores are greater than 100% (due to downward 'noise' in x_i). Using the terminology of the DEA literature, this efficiency metric allows for 'super-efficiency'.

We use the three output–input ratios as the clustering criteria: 'Energy transmission/Efficient cost' (r_1), 'Network length/Efficient cost' (r_2), and 'Customer number/Efficient cost' (r_3). Note that we take inefficiencies into account clustering by using the ratios $r_1, \ldots, r_3$ the efficient cost level characterized by the estimated cost frontier $C(\mathbf{y}_i)$. Assuming that the ratios $r_1, \ldots, r_3$ are normally distributed, we apply NMM for clustering. Denoting $\boldsymbol{V} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$, $|V| = \prod_{v=1}^{p} \sigma_v^2$ and p is the dimension of the observations, the probability density functions are defined as

$$f_i(\mathbf{r}_j;\theta_i) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{V}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{r}_j - \mu_i)^T \boldsymbol{V}^{-1}(\mathbf{r}_j - \mu_i)\right) \tag{27}$$

where

$$\hat{\mu}_i^{(m+1)} = \sum_{j=1}^{N} \tau_{ji}^{(m)} \mathbf{r}_j / \sum_{j=1}^{N} \tau_{ji}^{(m)} \tag{28}$$

$$\hat{V}_i^{(m+1)} = \sum_{j=1}^{N} \tau_{ji}^{(m)} (\mathbf{r}_j - \mu_i^{(m+1)})(\mathbf{r}_j - \mu_i^{(m+1)})^T / \sum_{j=1}^{N} \tau_{ji}^{(m)} \tag{29}$$

$$\hat{\pi}_i^{(m+1)} = \sum_{j=1}^{N} \tau_{ji}^{(m)} / N \tag{30}$$

$$\tau_{ji}^{(m)} = \frac{\pi_i^{(m)} f_i(\mathbf{r}_j;\theta_i^{(m)})}{\sum_{i=1}^{g} \pi_i^{(m)} f_i(\mathbf{r}_j;\theta_i^{(m)})} \tag{31}$$

The model parameters are estimated iteratively over (28)–(31) (see [41] for details). BIC is used for model selection as defined below,

$$\mathrm{BIC} = -2 \log L(\hat{\theta}) + d \log(pN), \tag{32}$$

where d represents the number of free parameters.

Table 1
Descriptive statistics of the input, output, and contextual variables of the empirical data. 'MEAN', 'STD', 'MIN', 'MAX', 'KURT', 'SKEW' represent the 'Mean', 'Standard deviation', 'Minimum value', 'Maximum value', 'Kurtosis' and 'Skewness', respectively. The data are averaged over a six-year period 2005–2010.

Variable	MEAN	STD	MIN	MAX	KURT	SKEW
x=total cost (1000€)	5052	10 144	139	64 326	22	4
y_1=energy transmission (GWh)	512	1026.65	15	6978	22	4
y_2=network length (km)	4370	10 465.63	46	68 349	26	5
y_3=customer number	37 650	73 856.08	24	426 769	16	4
z=underground cable proportion	0.23	0.28	0	1	0.43	1.27

3.2. Benchmarks obtained with StoNED and NMM

In this application, the NMM algorithm automatically identified four clusters, which match our prior classification of DMUs as rural, suburban, urban and industrial network firms. Table 2 reports the benchmark DMUs identified for each group using the clustering framework described above. For the largest clusters 1, 2, and 3, we can find several absolute benchmarks with efficiency scores greater than 100%. In these clusters, inefficient DMUs can choose one or more super-efficient DMUs to serve as the benchmark(s) based on such criteria as the similarity of activity or output structure, geographic location, or the aspired target level of efficiency. For the smallest cluster 4, interpreted as the group of industrial networks, we can only identify a relative benchmark, as none of the DMUs in this cluster currently operate with 100% efficiency. As we apply separate methods for efficiency analysis and clustering, it is possible to identify clusters in which all DMUs are inefficient.

In Tables 2 and 3 we label the four clusters identified by NMM as rural, suburban, urban and industrial network firms. This interpretation is justified by Table 3, where we report some descriptive statistics of the ratios $r_1,\ldots,r_3$ used as inputs to the NMM algorithm and the ratio of energy transmission to the network length. We interpret cluster 1 as the group of rural networks, because the network length is the main cost driver in the sparsely populated rural areas, while the number of customers and energy consumption are relatively small. In contrast, cluster 3 is interpreted as the group of urban networks as it has the highest number of customers relative to the efficient cost. The characteristics of cluster 2 are generally somewhere between those of clusters 1 and 3, so we interpret it as the group of suburban networks. Finally, cluster 4 is identified as the group of industrial networks as the DMUs in this group have notably higher energy transmission relative to the network length.

The four clusters are graphically illustrated in the three-dimensional output space (Fig. 1). The output variables can be scaled by the efficient cost level because the estimated cost function exhibits CRS. The observed DMUs belonging to different clusters are marked by different symbols. Benchmark DMUs are indicated by filled symbols and the empty ones represent inefficient DMUs. Note that the benchmarks are located furthest away from the viewer. Fig. 1 helps us to visualize the four groups identified by NMM.

Interestingly, we note that the clusters closely follow the ranking of DMUs according to the ratio of energy transmission to

Table 2
The cluster-specific benchmarking for each group. 'No. of DMUs' represent the number of DMUs in a given group, which all could choose references from the given benchmarks in the group.

Cluster No.	Benchmark	Efficiency (%)	No. of DMUs
Cluster 1 (rural)	Koillis-Satakunnan Sähkö Oy	108	26
	Järvi-Suomen Energia Oy	107	
	Lankosken Sähkö Oy	106	
	Sallila Sähkönsiirto Oy	105	
	Kokemäen Sähkö Oy	103	
	Tornionlaakson Sähkö Oy	101	
Cluster 2 (suburban)	Oulun Seudun Sähkö Verkkopalvelut Oy	119	33
	Herrfors Nät-Verkko Oy Ab	114	
	Paneliankosken Voima Oy	111	
	Tunturiverkko Oy	105	
	Pellon Sähkö Oy	101	
Cluster 3 (urban)	Oulun Energia Siirto ja Jakelu Oy	109	24
	Jakobstads Energiverk	108	
	Vantaan Energia Sähköverkot Oy	101	
Cluster 4 (industry)	Karhu Voima Oy	84	2

Table 3
Descriptive statistics of clustering based benchmarking. $r_1,\ldots,r_3$ represent the three output–input ratios for the cost frontier model.

	Cluster 1 (rural)	Cluster 2 (suburban)	Cluster 3 (urban)	Cluster 4 (industry)
Mean				
r_1=Energy transmission/efficient cost	0.075	0.124	0.158	0.156
r_2=Network length/efficient cost	1.403	1.110	0.550	0.117
r_3=Customer number/efficient cost	5.961	8.219	11.846	0.265
Energy transmission/network length	0.054	0.114	0.314	1.583
Standard deviation				
r_1=energy transmission/efficient cost	0.016	0.026	0.020	0.012
r_2=Network length/efficient cost	0.083	0.131	0.153	0.043
r_3=Customer number/efficient cost	0.977	1.776	2.281	0.228
Energy transmission/network length	0.014	0.026	0.108	0.686
Minimum				
r_1=Energy transmission/efficient cost	0.038	0.096	0.120	0.144
r_2=Network length/efficient cost	1.226	0.785	0.221	0.074
r_3=Customer number/efficient cost	3.651	3.763	8.009	0.038
Energy transmission/eetwork length	0.023	0.079	0.177	0.897
Maximum				
r_1=Energy transmission/efficient cost	0.102	0.168	0.210	0.168
r_2=Network length/efficient cost	1.611	1.374	0.801	0.161
r_3=Customer number/efficient cost	8.410	11.954	18.491	0.493
Energy transmission/network length	0.078	0.169	0.612	2.269

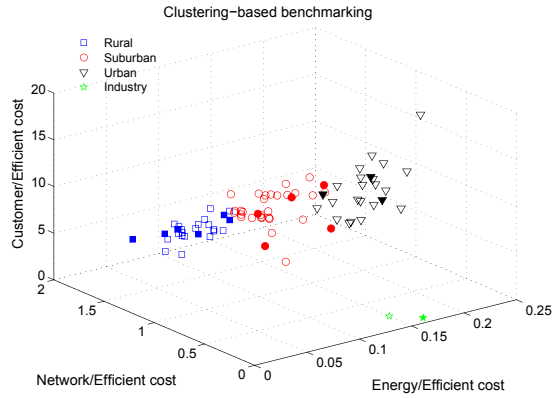


Fig. 1. Clustering based benchmarking results. The filled symbols represent the benchmarks for each corresponding cluster.

network length. Indeed, this ratio is commonly used as a simple indicator for distinguishing rural, suburban, and urban firms. Further, we note from Table 3 that the standard deviation of r_3 is relatively high in all clusters, which implies there is a lot of heterogeneity in the number of customers per efficient cost within each cluster. In other words, the energy transmission and the network length are the main outputs that distinguish the clusters. However, all three output variables are needed to identify these clusters: we cannot obtain these clusters by simply ranking DMUs according to any single output variable alone.

3.3. Comparison with DEA benchmarks

For comparison, the benchmarks obtained by DEA (CRS) are reported in Table 4. DEA identifies only four DMUs as 100% efficient. In other words, the DEA frontier is spanned by the four DMUs listed in Table 2. Virtually all previous DEA approaches to benchmarking identify some subset of these four DMUs as benchmarks for the inefficient units (one of the few exceptions is the context dependent DEA discussed in [52]). The weights assigned to the benchmark DMUs depend on the choice of orientation and the efficiency metric (e.g., radial, non-radial, or slack based), and the possible use of value judgments or preference information to impose weight restrictions. However, whichever efficiency metric or weight restrictions are used, DEA approaches identify some subset of these four DMUs as benchmarks in this application.

Comparing the four DEA benchmarks with those obtained with the combination of StoNED and NMM (reported in Tables 2 and 4), we find that all DEA efficient benchmarks are among the super-efficient DMUs according to StoNED estimation. In addition, the four DEA efficient benchmarks are included in clusters 1, 2, and 3 identified using the combination of StoNED and NMM. The results of DEA and StoNED support each other in these respects. In general, efficient units according to DEA are also efficient according to StoNED. However, the DEA efficient units are not necessarily the most efficient ones according to StoNED, as the comparison of Tables 2 and 4 illustrates.

The comparison of Tables 2 and 4 also reveals that the proposed combination of StoNED and NMM can identify a larger number of efficient benchmarks in each cluster. This is because the StoNED frontier does not envelop all DMUs, but allows for super-efficient units to be located above the efficient frontier. Further, the StoNED efficiency estimates provide a continuous ranking of efficient DMUs. In the present application we used the distance to frontier as efficiency metric, allowing for super-efficiency. Alternatively, one could apply the conditional expected value to transform the distance metric to efficiency scores restricted to the interval $[0, 1]$ (as discussed in [14]). The continuous rankings apply also in the latter case, but all DMUs are classified as inefficient (with efficiency score less than 100%) by construction.

More detailed DMU-specific benchmarks obtained with DEA and the combination of StoNED and NMM are provided as supplementary material to this article, available online at ⟨http://www.sciencedirect.com⟩.

Table 4
DEA benchmarking for each group. 'No. of DMUs' represents the number of DMUs in a given group.

	Benchmark	Efficiency (%)	No. of DMUs
Group 1	Lankosken Sähkö Oy	100	36
Group 2	Oulun Seudun Sähkö Verkkopalvelut Oy	100	25
Group 3	Oulun Energia Siirto ja Jakelu Oy	100	15
Group 4	Herrfors Nät-Verkko Oy Ab	100	9

4. Conclusions

In this paper, we presented a clustering based benchmarking framework to take into account the heterogeneity of firms and their operating environment, which ensures the long-term achievability of the targets for each DMU. In other words, the targets set for each firm are realistic given their similarities in, e.g., product, customer, and operation. The novelty of this framework lies not only in adjusting the benchmarking according to the intrinsic characteristics of the DMUs but also in its high flexibility due to the independence of the two stages, i.e., clustering and productive efficiency analysis, which can be tuned or optimized, separately, based on the customer needs or preferences. In particular, the inputs of the clustering and efficiency estimation are user-defined. Depending on the context according to which the benchmarking is expected, measure-specific clustering can be carried out by using a set of specific inputs or incorporating prior information. The efficiencies can be computed using different frontier models and the inputs can be customized depending on the factors users wish to evaluate. Also, the algorithms at each step could be freely chosen, modified or developed to meet the customer needs, allowing more freedom and better chance of getting the optimal targets. Further, the principle for choosing the frontier at each cluster allows multiple absolute benchmarks, forming a target pool for each DMU to choose from. In cases where no DMU achieves 100% efficiency, it ensures at least one reference for each user by outputting the relative benchmark, which is achievable at least in the long run since it considers the cluster-wise difference. Moreover, different clustering algorithms may provide different segmentations (methods such as fuzzy c-means allows one element belong to multiple groups), and multiple efficiency estimation methods can be combined into the proposed framework, both enlarging the pool size of the benchmarking for each DMU.

We applied the proposed cluster-specific framework to the Finland electricity distribution network data set, and the results are compared with those obtained from DEA. The clustering based method is shown to be able to well characterize each group under interest. Also, compared with DEA, more references are provided for each DMU, and targets with higher efficiencies could be identified using the proposed framework. Finally, the advantage of considering cluster-wise difference is well demonstrated by the concept of 'relative benchmark' which, otherwise, would lead to unrealistic benchmarking as shown by the references of the 4th group in the DEA outputs.

We believe that the flexible nature of the proposed approach is an attractive feature for practitioners who are free to choose the most suitable combination of efficiency assessment and clustering methods to match the objectives, information needs, and data availability in specific applications. From the academic point of view, the general nature of the proposed approach also poses an interesting research challenge: what is the optimal configuration and specification of the efficiency analysis and clustering methods when used in combination for benchmarking purposes? This question could be investigated by means of Monte Carlo simulations, which we suggest as an interesting avenue for future research.

Acknowledgments

We would like to thank Mr. Abolfazl Keshvari and three anonymous reviewers of this journal for their helpful comments.

Appendix A. Clustering criteria and distance measurements in hierarchical clustering

There are two types of hierarchical clustering algorithms, namely the agglomerative method and the divisive method, which

recursively combines or splits a set of objects into bigger or smaller groups based on a certain criterion. Commonly applied criteria include single linkage, complete linkage, average linkage, group average linkage and Ward's linkage, which are shown in (A.1)–(A.5). Notice in these equations that $D(G_i, G_j)$ and $d(\mathbf{r}_a, \mathbf{r}_b)$ each represents the distance between two clusters (G_i and G_j) and two firms ($\mathbf{r}_a$ and $\mathbf{r}_b$, $\mathbf{r}_a \in G_i, \mathbf{r}_b \in G_j$), respectively. The number of firms within groups G_i or G_j is shown as N_{G_i} or N_{G_j}, and *ESS* is the abbreviation of 'error sum of squares'.

$$D(G_i, G_j) = \min_{\mathbf{r}_a \in G_i, \mathbf{r}_b \in G_j} d(\mathbf{r}_a, \mathbf{r}_b) \tag{A.1}$$

$$D(G_i, G_j) = \max_{\mathbf{r}_a \in G_i, \mathbf{r}_b \in G_j} d(\mathbf{r}_a, \mathbf{r}_b) \tag{A.2}$$

$$D(G_i, G_j) = \frac{\sum_{a=1}^{N_{G_i}} \sum_{b=1}^{N_{G_j}} d(\mathbf{r}_a, \mathbf{r}_b)}{N_{G_i} \times N_{G_j}} \tag{A.3}$$

$$D(G_i, G_j) = d\left(\frac{\sum_{a=1}^{N_{G_i}} \mathbf{r}_a}{N_{G_i}}, \frac{\sum_{b=1}^{N_{G_j}} \mathbf{r}_b}{N_{G_j}} \right) \tag{A.4}$$

$$D(G_i, G_j) = ESS(G_i G_j) - ESS(G_i) - ESS(G_j) \quad \text{where}$$

$$ESS(G_i) = \sum_{a=1}^{N_{G_i}} \left| \mathbf{r}_a - \frac{1}{N_{G_i}} \sum_{w=1}^{N_{G_i}} \mathbf{r}_w \right|^2 \tag{A.5}$$

The group similarity is often scaled by distance, for which different measurements can be employed depending on the purpose and the characteristics of the firms. Among others, Euclidean distance, Mahalanobis distance, Manhattan distance, and Hamming distance are most commonly seen. These distances can be computed from (A.6) to A.9), respectively, where p ($p \in \{1, \infty\}$) is the dimension of each observation and 'Cov' represents the covariance matrix of two objects (firms are represented as objects here).

$$d(\mathbf{r}_a, \mathbf{r}_b) = \sqrt{\sum_{w=1}^{p} (r_{aw} - r_{bw})^2} \tag{A.6}$$

$$d(\mathbf{r}_a, \mathbf{r}_b) = \sqrt{(\mathbf{r}_a - \mathbf{r}_b)^T \, \mathrm{Cov}^{-1} (\mathbf{r}_a - \mathbf{r}_b)} \tag{A.7}$$

$$d(\mathbf{r}_a, \mathbf{r}_b) = \sum_{w=1}^{p} |r_{aw} - r_{bw}| \tag{A.8}$$

$$d(\mathbf{r}_a, \mathbf{r}_b) = \sum_{w=1}^{p} \kappa_w, \quad \kappa_w = \begin{cases} 1 & \text{if } r_{aw} \neq r_{aw} \\ 0 & \text{if } r_{aw} = r_{bw} \end{cases} \tag{A.9}$$

Appendix B. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org.10.1016/j.omega.2013.05.007.

References

[1] Farrell MJ. The measurement of productive efficiency. Journal of the Royal Statistical Society 1957;120(3):253–282.
[2] Charnes A, Cooper WW, Rhodes E. Measuring the inefficiency of decision making units. European Journal of Operational Research 1978;2(6):429–444.
[3] Thanassoulis E, Portela MCAS, Despic O. DEA—the mathematical programming approach to efficiency analysis. In: The measurement of productive efficiency and productivity growth. Oxford University Press; 2008.
[4] Liu JS, Lu LYY, Lu W, Lin BJY. A survey of DEA applications. Omega 2013;41(5):893–902.
[5] Paradi JC, Zhu H. A survey on bank branch efficiency and performance research with data envelopment analysis. Omega 2013;41(1):61–79.
[6] Bell RA, Morey RC. The search for appropriate benchmarking partners: a macro approach and application to corporate travel management. Omega 1994;22(5):477–490.
[7] Cook WD, Seiford LM, Zhu J. Models for performance benchmarking: measuring the effect of e-business activities on banking performance. Omega 2004;32(4):313–322.
[8] Hinojosa MA, Mármol AM. Axial solutions for multiple objective linear problems. An application to target setting in DEA models with preferences. Omega 2011;39(2):159–167.
[9] Epure M, Kerstens K, Prior D. Technology-based total factor productivity and benchmarking: new proposals and an application. Omega 2011;39(6):608–619.
[10] Adler N, Liebert V, Yazhemsky E. Benchmarking airports from a managerial perspective. Omega 2013;41(2):442–458.
[11] Samoilenko S, Osei-Bryson K. Using data envelopment analysis (DEA) for monitoring efficiency-based performance of productivity-driven organizations: design and implementation of a decision support system. Omega 2013;41(1):131–142.
[12] Aigner DJ, Lovell CAK, Schmidt P. Formulation and estimation of stochastic frontier models. J. Econometrics 1977;6:21–37.
[13] Meeusen W, van den Broeck J. Efficiency estimation from Cobb–Douglas production function with composed error. International Economic Review 1977;8:435–444.
[14] Kuosmanen T, Kortelainen M. Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. Journal of Productivity Analysis 2012;38(1):11–28.
[15] Kuosmanen T. Stochastic semi-nonparametric frontier estimation of electricity distribution networks: application of the StoNED method in the Finnish regulatory model. Energy Economics 2012;34:2189–2199.
[16] Johnson AL, Kuosmanen T. One-stage and two-stage DEA estimation of the effects of contextual variables. European Journal of Operational Research 2012;220(2):559–570.
[17] Johnson AL, Kuosmanen T. One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNEZD method. Journal of Productivity Analysis 2011;36(2):219–230.
[18] Orea L, Kumbhakar SC. Efficiency measurement using a latent class stochastic frontier model. Empirical Economics 2004;29(1):169–183.
[19] O'Donnell CJ, Rao DSP, Battese GE. Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. Empirical Economics 2008;34:231–255.
[20] Po R, Guh Y, Yang M. A new clustering approach using data envelopment analysis. European Journal of Operational Research 2009:276–284.
[21] Triantis K, Sarayia D, Seaver B. Using multivariate methods to incorporate environmental variables for local and global efficiency performance analysis. INFOR: Information Systems and Operational Research 2010;48(1/2):39–52.
[22] Fallah-Fini S, Triantis K, de la Garza JM, Seaver W. Measuring the efficiency of highway maintenance contracting strategies: a bootstrapped non-parametric meta-frontier approach. European Journal of Operational Research 2012;219(1):134–145.
[23] Banker RD. Maximum likelihood, consistency and data envelopment analysis: a statistical foundation. Management Science 1993;39(10):1265–1273.
[24] Kuosmanen T, Johnson AL. Data envelopment analysis as nonparametric least-squares regression. Operations Research 2010:149–160.
[25] Aparicio J, Ruiz JL, Sirvent I. Closest targets and minimum distance to the Pareto-efficient frontier in DEA. Journal of Productivity Analysis 2007;28:209–218.
[26] Aparicio J, Pastor JT. On how to properly calculate the Euclidean distance-based measure in DEA. Optimization 2012 http://dx.doi.org/10.1080/02331934.2012.655692.
[27] Johnson SC. Hierarchical clustering schemes. Psychometrika 1967;32(3):241–254.
[28] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York, USA: Springer Science and Business Media; 2009.
[29] Ward JH. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 1963;58(301):234–244.
[30] Mount DW. Bioinformatics: sequence and genome analysis. 2nd ed. New York: John Inglis; 2004.
[31] McGarigal K, Cushman S, Stafford SG. Multivariate statistics for wildlife and ecology research. New York: Springer-Verlag; 2000.
[32] Causton HC, Quackenbush J, Brazma A. Microarray/gene expressions data analysis: a beginner's guide. UK: Blackwell Science; 2003.
[33] Russell D. The principles of computer networking. Cambridge, UK: Cambridge University Press; 1989.
[34] Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. In: KDD-2000 workshop on text mining, 2000.
[35] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 2002;97:611–631.
[36] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, University of California Press; 1967. p. 281–97.
[37] Pelleg D, Moore A. X-means: extending K-means with efficient estimation of the number of clusters. In: ICML 2000. Morgan Kaufmann; 2000. p. 727–34.
[38] Melin P, Castillo O. Hybrid intelligent systems for pattern recognition using soft computing: an evolutionary approach for neural networks and fuzzy systems. Berlin, Germany: Springer-Verlag; 2005.

[39] Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. Genome Research 1999;9(11):1106–1115.
[40] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. New York: Wiley; 1990.
[41] McLachlan GJ, Peel D. Finite mixture models. New York, USA: John Wiley and Sons; 2000.
[42] Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics 2002;18(9):1194–1206.
[43] Vaithyanathan S, Dom B. Model-based hierarchical clustering. In: IPDPS'03. Morgan Kaufmann Publishers; 2003. p. 599–608.
[44] Zhong S, Ghosh J. A unified framework for model- based clustering. Journal of Machine Learning Research 2003;4:1001–1037.
[45] Smyth P. Model selection for probabilistic clustering using cross-validated likelihood. Statistics and Computing 2000;9(1):63–72.
[46] Ji Y, Wu C, Liu P, Wang J, Coombes RK. Applications of beta-mixture models in bioinformatics. Bioinformatics 2005;21(9):2118–2122.
[47] Akaike H. A new look at the statistical identification model. IEEE Transactions on Automatic Control 1974;19:716–723.
[48] Schwarz G. Estimating the dimension of a model. Annals of Statistics 1978;6(2):461–464.
[49] Bozdogan H. Model selection and Akaike I nformation Criterion (AIC): the general theory and its analytic extensions. Psychometrika 1987;52(3):345–370.
[50] Biernacki C, Govaert G. Choosing models in model-based clustering and discriminant analysis. Journal of Statistical Computation and Simulation 1999;64(1):49–71.
[51] Pan W. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. Bioinformatics 2006;22(7):795–801.
[52] Seiford LM, Zhu J. Context-dependent data envelopment analysis: measuring attractiveness and progress. Omega 2003;31:397–408.

Publication IV

Xiaofeng Dai. NMM-StoNED: a normal mixture model based stochastic semi-parametric benchmarking method. *International Journal of Business and Management Study*, 2 (2), 2015.

© 2015 Canadian Center of Science and Education Publications.
Reprinted with permission.

NMM-StoNED: a normal mixture model based stochastic semi-parametric benchmarking method

Xiaofeng Dai

***Abstract*—This paper presents a novel benchmarking tool, NMM-StoNED, which identifies the best practices closely located with each decision making unit (DMU) in the input-output space. Unlike the conventional techniques such as DEA where the success recepies of the benchmarks may not be transferable to all DMUs given their differences in, e.g., the operational scales, best practices identified by this method do not suffer from these problems and offer more practical values. NMM-StoNED is a specific configuration of the clustering and efficiency estimation algorithms in the benchmarking framework previously presented. This combination is able to cluster DMUs into less ambiguous groups and model the inefficiencies in a stochastic semi-nonparametric framework, which produces more accurate results than conventional benchmarking techniques such as DEA or other combinations such as the integration of K-means and StoNED. The performance comparison between NMM-StoNED and DEA has previously been reported, and the superiorities of StoNED over other productive efficiency analysis methods have been thoroughly investigated. Here we focus on showing the advantages of NMM in the clustering based benchmarking framework, for which, an empirical study using the Finland energy regulation data was conducted. This study contributes in its systematic evaluations on the performance of NMM-StoNED under various conditions which provide solid specifications on this algorithm, availing its practical use.**

***Keywords*—benchmarking, normal mixture model (NMM), data envelopment analysis (DEA), stochastic semi-nonparametric envelopment of data (StoNED)**

I. Introduction

Benchmarking, the process of comparing the performance of one decision making unit (DMU) against that of the DMUs with the 'best practice', has multiple applications, including offering the general insight of a given business sector, facilitating the manager on decision making, and providing the backbone of incentive provision for the regulators in the context of multiple agents [1]. DEA (data envelopment analysis) is conventionally applied in benchmarking, where the intensity weights strictly positive from the frontier estimation are considered as the best practices in benchmarking [2]. However, the success formula of the benchmarks identified may not be transferrable to a given DMU if they differ greatly on their, e.g., input and output structure. Also, as a deterministic method geared towards efficiency estimation, DEA doesn't take consider the stochasticity in its modelling framework. Thus, DEA is sensitive to both the heterogeneity and random noise of the DMUs in benchmarking.

Xiaofeng Dai
JiangNan University
China P. R.
Email: xiaofeng.dai@me.com

We have proposed a clustering based benchmarking framework in [4], where it segments the DMUs into groups based on user-specified metrics (e.g., the input-output vectors or their projections on the estimated frontier) using a clustering technique, and the benchmark(s) are identified according to the efficiency scores estimated from productive efficiency analysis within each cluster. We have shown that such a framework is flexible in choosing the clustering and efficiency analysis algorithms and these problems could be efficiently solved if the method at each step is appropriately selected. However, for what combination this method achieves the best performance is still left for discussion.

Typical clustering approaches can be classified into three categories, i.e., the hierarchical methods, the partitioning methods, and the model-based methods [5]. Hierarchical algorithms recursively combines or splits a set of objects into bigger or smaller groups based on a certain distance measurement and stops when meeting a certain criterion [6]. Methods of this class are conceptually intuitive and computationally simple which, however, could not determine the number of groups automatically, needs expert domain knowledge to define the distance measurement and is problem-specific. Partitioning methods iteratively reallocate data points across groups until no further improvement is obtainable [5], [11], with K-means being the most representative algorithm of this class [11]. Partitioning methods are widely used due to their computational simplicity and nonparametric structure which, however, needs pre-specification of the number of clusters. Model-based techniques optimise the fitness between the data and the model where the data is assumed to be generated [14]. Model based methods are superior over other methods in their automatic determination of the number of clusters, robustness to outliers, and probabilistic nature [5]. Among others, NMM (normal mixture model) is the most widely applied method of this class since normal distribution is the most commonly encountered distribution in practice.

Traditional productive efficiency analysis methods can be grouped based on two properties, i.e., parametric or non-parametric, and deterministic or stochastic. Many statistical methods can be used for productive efficiency analysis, with the most widely applied being DEA and SFA (stochastic frontier analysis), where DEA is non-parametric but deterministic and SFA is stochastic but parametric [8]. StoNED (stochastic semi-nonparametric envelopment of data) is a recently developed technique that melds the merits of DEA and SFA where the inefficiencies are estimated in a stochastic semi-nonparametric fashion. Unlike the semiparametric variate of

SFA, StoNED builds directly on the axioms of the production theory such as free disposability and convexity instead of making any assumptions on the functional form or smoothness [9]. On the other hand, StoNED uses information of all observations in the data set to estimate the frontier rather than a few influential ones as adopted by DEA, making it less sensitive to outliers than DEA besides its insensitivity to the random noise.

Given the advantages of NMM and StoNED in clustering and efficiency estimation, respectively, we are motivated to fit these two algorithms in the clustering based benchmarking framework presented in [4]. This method, named NMM-StoNED here, detects the heterogeneous structure of the data, groups similar DMUs into unambiguous clusters, and ranks them within each cluster by the estimated efficiencies according to which the best practice is identified for each group. The superiorities of NMM-StoNED over DEA have been demonstrated in [4] using Finland energy regulation data from EMA (Energy Market Authority), and the advantages of StoNED over other efficiency analysis methods such as DEA and SFA have been studied in [10]. Here we focus on evaluating the performance of combining NMM with StoNED as compared with integrating other clustering techniques with StoNED in benchmarking. For this, we compared NMM with K-means, the most widely applied clustering technique due to its simple yet powerful features, in this clustering based benchmarking framework with an empirical study.

The rest of paper is organized as 'Method', 'Empirical study' and 'Conclusion'. The technical details of NMM and StoNED are described in the 'Method' section. In the 'Empirical study', the 'Data and methods' and 'Results and discussion' are described by sub-sections. The 'Conclusion' section finalizes this paper by summarizing the work and main contributions, and pointing out the future direction.

II. Method

The proposed method, NMM-StoNED, combines the NMM and StoNED into a unified framework. One can either measure the efficiencies of all DMUs using the whole data set before clustering, or compute the efficiencies using segment frontier after clustering if the number of DMUs in each cluster is sufficiently large [4]. The first alternative was used here given the limited size of our empirical data. The estimation process comprises of 1) estimating the efficiencies of all DMUs from the whole data set using StoNED; and 2) clustering DMUs using NMM and identifing the best practices in each group.

A. Efficiency estimation using StoNED

Given the standard multiple-input $\mathbf{r}_i$, single-output y_i, cross-sectional productive efficiency analysis model $y_i = f(\mathbf{r}_i) - u_i + v_i, \forall i = 1, \ldots, N.$ where f satisfies monotonicity and concavity, $u_i > 0$ is an asymmetric inefficiency term and v_i is a stochastic noise term, StoNED uses a two-stage strategy in efficiency estimation [9]. In Stage 1, the shape of the function f is estimated by convex nonparametric least squares (CNLS) regression. In Stage 2, the inefficiency u is computed from the variances (σ_u^2, σ_v^2), which are estimated based on the skewness of the CNLS residuals (obtained from Stage 1) using, e.g., the method of moments. In the second stage, additional distributional assumptions are typically assumed, including, e.g., the asymmetric distribution for u_i with positive mean μ and finite variance σ_u^2, and a symmetric distribution for v_i with zero mean and constant finite variance σ_v^2.

Mathematically, the first stage is equivalent to (1) to (4) [9],

$$\min_{\boldsymbol{v},\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{i=1}^{n} \varepsilon_i^2 \quad \text{such that} \tag{1}$$

$$y_i = \alpha_i + \boldsymbol{\beta}_i^{'}\mathbf{r}_i + \varepsilon_i \tag{2}$$

$$\alpha_i + \boldsymbol{\beta}_i^{'}\mathbf{r}_i \leq \alpha_h + \boldsymbol{\beta}_h^{'}\mathbf{r}_i, \forall h, i = 1 \ldots n \tag{3}$$

$$\boldsymbol{\beta}_i \geq \mathbf{0}. \forall i = 1 \ldots n \tag{4}$$

where α_i and $\boldsymbol{\beta}_i$ are coefficients specific to observation i and v_i captures its random noise. In Stage 2, the inefficiency is computed using the distribution of the CNLS residuals $\hat{\varepsilon}_i$ (note that $\varepsilon = v_i + u_i$). Assuming that the inefficiency and noise terms follow the half-normal and normal distributions, respectively, the 2^{nd} and 3^{rd} central moments of the composite error distribution are

$$M_2 = [\frac{\pi - 2}{\pi}]\sigma_u^2 + \sigma_v^2, \quad M_3 = -(\sqrt{\frac{2}{\pi}})[\frac{4}{\pi} - 1]\sigma_u^3, \tag{5}$$

which can be estimated using the CNLS residuals

$$\hat{M}_2 = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \overline{\varepsilon})^2/n, \quad \hat{M}_3 = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \overline{\varepsilon})^3/n. \tag{6}$$

The standard deviations of the inefficiency and error term are then computed from

$$\hat{\sigma}_u = \sqrt[3]{\frac{\hat{M}_3}{(\sqrt{\frac{2}{\pi}})[\frac{4}{\pi} - 1]}}, \quad \hat{\sigma}_v = \sqrt{\hat{M}_2 - [\frac{\pi - 2}{\pi}]\hat{\sigma}_u^2}. \tag{7}$$

The conditional distribution of the inefficiency u_i given ε_i is a zero-truncated normal distribution with mean $\mu_\star = -\varepsilon_i\sigma_u^2/(\sigma_u^2 + \sigma_v^2)$ and variance $\sigma_\star^2 = \sigma_u^2\sigma_v^2/(\sigma_u^2 + \sigma_v^2)$. Let ϕ and Φ represent the standard normal density function and the standard normal cumulative distribution function, respectively, the inefficiencies are computed by $E(u_i|\varepsilon_i) = \mu_\star + \sigma_\star[\frac{\phi(-\mu_\star/\sigma_\star))}{1-\Phi(-\mu_\star/\sigma_\star)}]$.

B. Cluster-specific benchmark identification using NMM

In this step, the metrics dominating the heterogeneity of the data and following (or convertible to) normal distribution were specified and used as the input of NMM. If the input does not follow normal distribution or is a composite of multiple distributions, the mixture model of the corresponding distribution or a joint mixture model [3] would be required.

Assume that each observation $\mathbf{r}$ is drawn from g mixed normal distributions where, for each normal distribution f_i, it has the prior probability π_i and parameters θ_i, NMM optimises the fitness between the data and model $f(\mathbf{r}; \Theta) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{r}; \boldsymbol{\theta}_i)$. Note that $\Theta = \{(\pi_i, \boldsymbol{\theta}_i) : i = 1, \ldots, g\}$ denotes all unknown parameters, $0 \leq \pi_i \leq 1$ for any i and $\sum_{i=1}^{g} \pi_i = 1$. Expectation Maximization (EM) algorithm is used to iteratively estimate the parameters by maximising the data log-likelihood $\log L(\Theta) = \sum_{j=1}^{N} \log([\sum_{i=1}^{g} \pi_i f_i(\mathbf{r}_j; \boldsymbol{\theta}_i)])$, where $R = \{\mathbf{r}_j : j = 1, \ldots, N\}$ and N is the total number of observations. The problem is casted in the framework of incomplete data using a

dummy variable I_{ji} to indicate whether $\mathbf{r}_j$ comes from component i. Thus, $\log L_c(\Theta) = \sum_{j=1}^{N}\sum_{i=1}^{g} I_{ji}\log\left(\pi_i f_i(\mathbf{r}_j;\boldsymbol{\theta}_i)\right)$. At the m^{th} iteration of the EM algorithm, the E (expectation) step computes the expectation of the complete data log-likelihood Q

$$\begin{aligned} Q(\Theta;\Theta^{(m)}) &= E_{\Theta^{(m)}}(\log L_c|R) \\ &= \sum_{j=1}^{N}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log\left(\pi_i f_i(I_j;\boldsymbol{\theta}_i)\right), \end{aligned} \tag{8}$$

and the M (maximisation) step updates the parameter estimates to maximize Q. The algorithm is iterated until convergence. Note that I's are replaced with τ's in (8), where $\tau_{ji} = E[I_{ji}|\mathbf{r}_j, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_g; \hat{\pi}_1, ..., \hat{\pi}_g]$. The set of parameter estimates $\left\{\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_g; \hat{\pi}_1, ..., \hat{\pi}_g\right\}$ is a maximizer of the expected log-likelihood for given τ_{ji}'s, and each $\mathbf{r}_j$ is assigned to its component by $\{i_0|\tau_{ji_0} = \max_i \tau_{ji}\}$. In NMM, the probability density function of f_i is defined as $f_i(\mathbf{r}_j;\theta_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{V}|^{\frac{1}{2}}}\exp\big(-\frac{1}{2}(\mathbf{r}_j-\mu_i)^T\boldsymbol{V}^{-1}(\mathbf{r}_j-\mu_i)\big)$. Note that $\boldsymbol{V} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$, $|V| = \prod_{v=1}^{p}\sigma_v^2$ and p is the dimension of the observations, whose parameters are estimated iteratively over the following equations [12].

$$\hat{\mu}_i^{(m+1)} = \sum_{j=1}^{N}\tau_{ji}^{(m)}\mathbf{r}_j \Big/ \sum_{j=1}^{N}\tau_{ji}^{(m)}$$

$$\hat{V}_i^{(m+1)} = \sum_{j=1}^{N}\tau_{ji}^{(m)}(\mathbf{r}_j-\mu_i^{(m+1)})(\mathbf{r}_j-\mu_i^{(m+1)})^T \Big/ \sum_{j=1}^{N}\tau_{ji}^{(m)}$$

$$\hat{\pi}_i^{(m+1)} = \sum_{j=1}^{N}\tau_{ji}^{(m)}/N \qquad \tau_{ji}^{(m)} = \frac{\pi_i^{(m)} f_i(\mathbf{r}_j;\theta_i^{(m)})}{\sum_{i=1}^{g}\pi_i^{(m)} f_i(\mathbf{r}_j;\theta_i^{(m)})}$$

Bayesian information criterion (BIC) [13], the most widely used model selection method, was used here to determine the best fitting model as well as the optimal number of clusters if not particularly specified BIC $= -2\log L(\hat{\theta}) + d\log(pN)$, where d represents the number of free parameters.

Once the DMUs are properly segregated, we rank the DMUs within each cluster by their efficiencies, and the best practice(s) within each cluster are considered the benchmarks of other units belonging to this group. As pointed out in [4], the 'best practice' may not achieve 100% efficiency, and is called the 'relative benchmark' to differentiate it from the 'absolute benchmark' which achieves, and more than one 'absolute benchmark' may exist for one group if multiple DMUs achieve 100% efficiency. Relative benchmark is defined as $h = \{i|\max_{i=1}^{N_{G_j}}\zeta_i\}, \max_{i=1}^{N_{G_j}}\zeta_i < 1$, and absolute benchmark is $h = \{i|\zeta_i \geq 1\}, \max_{i=1}^{N_{G_j}}\zeta_i \geq 1$, where h denotes the frontier, ζ_i represents the efficiency of DMU i ($i \in 1 \ldots N_{G_j}$ in group G_j), N is the number of DMUs, g is the number of groups identified, and G_j has N_{G_j} DMUs.

III. Empirical study

A. Data and methods

Our empirical data comes from the Energy Market Authority (EMA) website (www.emvi.fi), which consists of 85 electricity suppliers and are the six-year average over the period 2005-2010 [4], [7]. Recently, EMA has replaced the conventional DEA and SFA by StoNED after a rigorous evaluation process [7]. Also, provided with the advantages of StoNED in overcoming the pitfalls of DEA and SFA [10], we fitted StoNED in this framework, and focused on evaluating the performance of NMM in improving the accuracy of efficiency estimation when combined with StoNED. For the purpose of comparison, K-means, a simple yet powerful and most widely applied clustering technique, was chosen.

We used the cost frontier model, $x_i = C(\mathbf{y}_i)\cdot\exp(\delta z_i + u_i + v_i)$, as adopted by EMA [7], in this empirical study, where C denotes the frontier cost function. This model adds a contextual variable z and its weight δ to the conventional cost frontier model. The variable z is the proportion of the underground cables in the total network length which captures the heterogeneity of the electricity suppliers in Finland, since the underground cables are widely used in urban and suburban regions but not in rural areas. In this model, the total cost (x) is used as the single input, and three variables, i.e., 'Energy transmission' (GWh of 0.4 kV equivalents, y_1), 'Network length' (km, y_2), and 'Customer number' (y_3) are specified as the outputs ($\mathbf{y}$). We used the three output-input ratios from productive efficiency analysis as the input variables for clustering, i.e., 'Energy transmission/Efficient cost' (r_1), 'Network length/Efficient cost' (r_2), and 'Customer number/Efficient cost' (r_3), where the efficient cost is computed as the estimated cost frontier '$C(\mathbf{y}_i)$' to take into account the efficiencies in segmentation. In addition, the actual cost was used in the inputs, i.e., 'Energy transmission/Actual cost' (r_1), 'Network length/Actual cost' (r_2), and 'Customer number/Actual cost' (r_3), to exclude the influence of the efficiencies in the analysis as a comparison. Note that the efficient cost is computed as the actual cost multiplied by the firm efficiency. We used the descriptive statistics of the clustered groups to evaluate the clustering accuracy, assuming that better clustering results in more distant inter-group means, less cross-group overlaps and lower within-group standard deviations.

B. Results and discussion

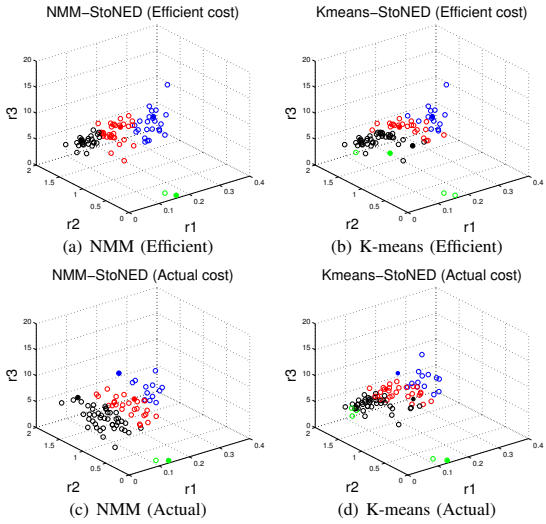
The 85 firms were grouped into four clusters, which consist of 26, 33, 24 and 2 DMUs, respectively, for clusters 1 to 4. The descriptive statistics, including mean, standard deviation and parameter ranges of $r_1 \ldots r_3$ and 'Energy transmission/Network length', are summarized for groups clustered by NMM and K-means in Table 1. Efficient and actual costs are used as the denominator of the inputs in the upper and lower panel of Table 1, respectively.

Let's first analyze the scenarios where efficient cost is used for computing the clustering inputs. It is seen that the groups clustered using NMM are characteristic of the four types of electricity networks in Finland, but with K-means the statistics are not as representative as such especially for the 4^{th} cluster (the industrial network). Specifically, the rural area consumes less energy than the other regions given its sparse population in Finland and there is no significant difference among suburban,

urban and industrial customers. This property is represented by r_1, and better captured in NMM-clustered groups than those clustered by K-means, since the distance between cluster 1 and the average of the other clusters is (0.124+0.158+0.156)/3-0.075=0.071 in NMM-clustered groups which is larger than that of K-means, i.e., (0.137+0.162+0.124)/3-0.095=0.046 (Table 1). The distance between the customer and electricity producer decreases from the rural to the industrial group, leading to a declining trend in the 'Network length' from clusters 1 to 4. This is well-captured by r_2 in NMM-clustered groups but is violated by the industrial cluster when the groups are clustered by K-means (i.e., the distance is 0.735 in the industrial group which is bigger than 0.529, the distance in the urban cluster). The number of customers increases from the rural to urban regions, and only a few industrial customers exist in Finland. This property is captured by r_3 in both NMM and K-means clustered groups. However, as the standard deviation of the group means is slightly larger in NMM-clustered groups than that in the K-means case, we'd say that groups are more clearly separated by NMM than K-means regarding this parameter. Here, we also examined the 'Energy transmission/Network length', since it merges r_1 and r_2 (the parameters that capture the principle differences between NMM and K-means in separating these groups given their statistics) and should represent the major distinction between the four groups as well as different clustering techniques. As seen from Table 1, the standard deviation of the group means is much larger in NMM-clustered groups (0.623) than that in the case of K-means (0.288), the average standard deviation of the groups is lower in case of NMM (0.209) compared with K-means (0.285), and there is no adjacent group overlap in NMM separated clusters but is 0.323 on average in the case of K-means. Thus, it is concluded that NMM could separate the four types of electricity suppliers into more appropriate groups compared with K-means in this empirical study.

The same conclusions can be drawn when the actual cost is used in the inputs as seen from Table 1. Thus, NMM performs better than K-means in this real case application regardless of whether the efficiencies are taken into account in computing the clustering inputs. However, using efficient cost in the inputs indeed groups the DMUs into more distant clusters than using the actual cost no matter whether NMM or K-means is used. For example, the averages of the standard deviation and overlapping range are lower in most cases when the efficient cost is used than those computed using the actual cost, indicating a higher within-group homogeneity and a larger inter-group distance when efficiencies are included in grouping. Also, the standard deviations of the group means are mostly larger when the efficient cost is used in the inputs than those computed using the actual cost, which again shows a larger inter-group distance among the four clusters.

The superiority of NMM over K-means in separating the rural, urban, suburban and industrial electricity networks in Finland is also illustrated in Figure 1. In this figure, each color represents one type of electricity supplier. There is a clear trend from the rural to urban areas (colored in black, red,



(a) NMM (Efficient) (b) K-means (Efficient)
(c) NMM (Actual) (d) K-means (Actual)

Figure 1. Comparison of NMM and K-means using EMA data. The efficient cost (a and b) and actual cost (c and d) are used in the inputs. The filled dots are the best performing unit for each cluster. 'Black', 'red', 'blue' and 'green' represent 'rural', 'suburban', 'urban' and 'industrial' networks, respectively.

and blue, respectively) along the three axes and the industrial cluster (shown in green) is distinctively separated from the other groups in NMM clustering, regardless of whether the efficiency is taken into account; yet when K-means is used, the boundaries become ambiguous especially for the industrial group where a few units are scattered into the rural cluster. More importantly, notice that the filled dots (representing the best performing DMU in a given cluster) may differ when different clustering techniques are used, resulting in different benchmarks for a given DMU. Take the industrial group as an example, its best performing unit is within the rural area in K-means clustering when efficient cost is used in the inputs which, once chosen as the benchmark for the industrial group, will become an unrealistic goal for this cluster given their large differences in, e.g., the input-output space.

IV. Conclusions

We present a combination of the NMM based clustering and the StoNED efficiency estimation technique in the benchmarking framework previously presented in [4]. It inherits the advantages of NMM such as automatic determination of the number of clusters and insensitivity to random noise, and the benefits of StoNED in its stochastic and semi-parametric modelling. With one empirical application we show that the DMUs could be clustered into groups having less ambiguous boundaries than other clustering techniques such as K-means. The superiorities of StoNED over other productive efficiency analysis methods such as DEA and SFA have been previously studied in [10]. Further, the benefits of combining

Efficient cost	NMM					K-means				
Mean	Cluster 1	Cluster 2	Cluster 3	Cluster 4	STD(Mean)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	STD(Mean)
r_1	0.075	0.124	0.158	0.156	0.034	0.095	0.137	0.162	0.124	0.024
r_2	1.403	1.110	0.550	0.117	0.497	1.273	0.911	0.529	0.735	0.273
r_3	5.961	8.219	11.846	0.265	4.204	6.504	9.785	13.227	1.986	4.149
ET/NL	0.054	0.114	0.314	1.583	0.623	0.080	0.174	0.341	0.827	0.288
STD					Mean(STD)					Mean(STD)
r_1	0.016	0.026	0.020	0.012	0.019	0.027	0.022	0.020	0.041	0.028
r_2	0.083	0.131	0.153	0.043	0.103	0.196	0.268	0.177	0.628	0.317
r_3	0.977	1.776	2.281	0.228	1.316	1.015	0.784	1.707	1.729	1.309
ET/NL	0.014	0.026	0.108	0.686	0.209	0.039	0.090	0.113	0.898	0.285
[min, max]					Mean(OL)					Mean(OL)
r_1	[0.038,0.102]	[0.096,0.168]	[0.120,0.210]	[0.144,0.168]	0.026	[0.038,0.168]	[0.084,0.174]	[0.135,0.210]	[0.059,0.168]	0.052
r_2	[1.226,1.611]	[0.785,1.374]	[0.221,0.801]	[0.074,0.161]	0.055	[0.642,1.611]	[0.357,1.374]	[0.221,0.923]	[0.074,1.506]	0.667
r_3	[3.651,8.410]	[3.763,11.954]	[8.009,18.491]	[0.038,0.493]	2.864	[4.552,8.085]	[8.297,11.302]	[11.612,18.491]	[0.038,3.763]	0
ET/NL	[0.023,0.078]	[0.079,0.169]	[0.177,0.612]	[0.897,2.269]	0	[0.023,0.237]	[0.064,0.489]	[0.153,0.612]	[0.039,2.269]	0.323

Actual cost	NMM					K-means				
Mean	Cluster 1	Cluster 2	Cluster 3	Cluster 4	STD(Mean)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	STD(Mean)
r_1	0.093	0.137	0.163	0.156	0.027	0.097	0.133	0.163	0.105	0.026
r_2	1.269	0.887	0.600	0.117	0.421	1.246	0.903	0.600	0.907	0.229
r_3	6.658	9.494	12.915	0.265	4.643	6.572	9.806	12.915	3.105	3.653
ET/NL	0.081	0.188	0.318	1.583	0.607	0.087	0.183	0.318	0.663	0.218
STD					Mean(STD)					Mean(STD)
r_1	0.027	0.023	0.021	0.012	0.021	0.030	0.023	0.021	0.043	0.029
r_2	0.228	0.289	0.238	0.043	0.200	0.233	0.306	0.238	0.647	0.356
r_3	1.549	1.885	2.278	0.228	1.485	1.307	1.600	2.278	2.437	1.906
ET/NL	0.045	0.109	0.128	0.686	0.242	0.050	0.111	0.128	0.867	0.289
[min, max]					Mean(OL)					Mean(OL)
r_1	[0.038,0.150]	[0.095,0.174]	[0.135,0.210]	[0.144,0.168]	0.053	[0.038,0.168]	[0.095,0.174]	[0.135,0.210]	[0.059,0.168]	0.048
r_2	[0.542,1.611]	[0.331,1.296]	[0.221,1.102]	[0.074,0.161]	0.508	[0.542,1.611]	[0.331,1.374]	[0.221,1.102]	[0.074,1.506]	0.828
r_3	[3.651,10.579]	[5.393,12.803]	[8.981,18.491]	[0.038,0.493]	3.003	[3.763,9.478]	[6.913,12.803]	[8.981,18.491]	[0.038,5.700]	2.129
ET/NL	[0.023,0.266]	[0.073,0.489]	[0.128,0.612]	[0.897,2.269]	0.185	[0.023,0.266]	[0.073,0.489]	[0.128,0.612]	[0.039,2.269]	0.346

Table 1. Descriptive statistics of groups clustered using efficient (upper panel) and actual (lower panel) costs in the inputs. ET/NL is Energy transmission/Network length. 'STD(Mean)' represents the standard deviation of the mean. 'Mean(STD)' is the average of the standard deviation. Overlap is computed between every adjacent 2 ranges, 'Mean(OL)' is the average length of 3 overlaps among 4 clusters.

NMM and StoNED as compared with the traditional DEA in benchmarking has been previously demonstrated by an empirical application in [4]. Thus, the performance of the proposed configuration in the clustering based benchmarking framework [4], i.e., NMM-StoNED, has been well-surrounded and is suggested to use if no specific needs to meet.

With the metrics selected as the input of clustering, we obtained four mutually exclusive clusters, each corresponds to a well-defined type of energy supplier. It is worth mentioning that with different metrics as the inputs, the clustering results may differ. Thus, one need to identify the principle statistics dominating the heterogeneity of the DMUs if not otherwise specified before clustering. If the input metrics do not follow or are not convertible to the normal distribution, a mixture model of the corresponding distribution or a joint mixture model [3] need to be used. Also, the computational complexity increases with the number of inputs. Therefore, techniques such as principle component analysis are needed to capture the main properties needed for clustering.

This paper successfully applies NMM-StoNED to energy regulation data which, however, is not restricted to such an area. It is applicable to any problems where the distribution of the evaluating metric is or convertible to normal distribution. Here we focus on applying NMM-StoNED in the cross-section setting, which could be used for panel data as well. To solve more practical benchmarking problems especially those that are problematic using conventional methods, more applications are worthwhile to explore.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (31471251) and the Fundamental Research Funds for the Central Universities (JUSRP11507). I thank Prof. Timo Kuosmanen for his insightful advice.

REFERENCES

[1] P. Bogetoft and K. Nielsen, "Internet based benchmarking", Group Decision and Negotiation, vol. 14, 2005, pp. 195-215.

[2] L. Botti, W. Briec and G. Cliquet, "Plural forms versus franchise and company-owned systems: a DEA approach of hotel chain performance", Omega, vol. 37, 2009, pp. 566-578.

[3] X. F. Dai, T. Erkkila, O. Yli-Harja and H. Lahdesmaki, "A joint finite mixture model for clustering genes from independent Gaussian and beta distributed data", BMC Bioinformatics, vol. 10, 2009, doi:10.1186/1471-2105-10-165.

[4] X. F. Dai and T. Kuosmanen, "Best-practice benchmarking using clustering methods: Application to energy regulation", Omega, vol. 42, 2013, pp. 179-188.

[5] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation", Journal of the American Statistical Association, vol. 97, 2002, pp. 611-631.

[6] S. C. Johnson, "Hierarchical clustering schemes", Psychometrika, vol. 32, 1967, pp. 241-254.

[7] T. Kuosmanen, "Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the finnish regulatory model", Energy Economics, vol. 34, 2012, pp. 2189-2199.

[8] T. Kuosmanen and A. L. Johnson, "Data envelopment analysis as nonparametric least-squares regression", Operations Research, vol. 58 , 2010, pp. 149-160.

[9] T. Kuosmanen and M. Kortelainen, "Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints", Journal of Productivity Analysis, vol. 38, 2012, pp. 11-28.

[10] T. Kuosmanen, A. Saastamoinen and T. Sipilainen, "What is the best practice for benchmark regulation of electricity distribution? comparison of DEA, SFA and StoNED methods", Energy Policy, vol. 61, 2013, pp. 740-750.

[11] J. MacQueen, "Some methods for classification and analysis of multivariate observations", In Proceedings of the 5th berkeley symposium on mathematical statistics and probability, 1967, pp. 281-297.

[12] G. J. McLachlan and D. Peel, "Finite mixture models". New York, USA: John Wiley & Sons, 2000.

[13] W. Pan, "Incorporating gene functions as priors in model-based clustering of microarray gene expression data", Bioinformatics, vol. 22, 2006, pp. 795-801.

[14] S. Zhong and J. Ghosh, "A unified framework for model-based clustering", Journal of Machine Learning Research, vol. 4, 2003, pp. 1001-1037.

This dissertation endeavors to explore the interdisciplinary applications of computational methods in quantitative economics. Particularly, this thesis focuses on problems in productive efficiency analysis and benchmarking that are hardly approachable or solvable using conventional methods. The methods developed benefit this field and the problem-solving perspectives lighten a new direction.

Aalto-DD 134/2016

ISBN 978-952-60-6906-7 (printed)
ISBN 978-952-60-6907-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Business
Department of Information and Service Economy
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS